Estadística aplicada a las Ciencias Sociales

La fiabilidad de los tests y escalas

©Pedro Morales Vallejo
Universidad Pontificia Comillas, Madrid
Facultad de Ciencias Humanas y Sociales
(última revisión, 18 de Septiembre de 2007)

Índice

1. Conceptos preliminares básicos 3
   1. *Equivocidad del concepto de fiabilidad*........ 3
   2. *Fiabilidad y precisión de la medida* 3
   3. *Fiabilidad y margen de error en al medida*........ 3
   4. *Fiabilidad y predictibilidad*........ 3
   5. *Fiabilidad y validez* 4
   6. *Fiabilidad y validez: errores sistemáticos y errores aleatorios*........ 4
   7. *La fiabilidad no es una característica de los instrumentos*........ 4
   8. *Fiabilidad y diferencias: teoría clásica de la fiabilidad* 5

2. Enfoques y métodos en el cálculo de la fiabilidad........ 5
   2.1· Método: Test-retest........ 5
   2.2. Método: Pruebas paralelas........ 6
   2.3. Método: Coeficientes de consistencia interna 6

3. Los coeficientes de *consistencia interna*
   concepto y fórmula básica de la fiabilidad........ 7

4. Requisitos para una fiabilidad alta 8

5. Las fórmulas Kuder -Richardson 20 y α de Cronbach 11

6. Factores que inciden en la *magnitud* del coeficiente de fiabilidad 13

7. Interpretación de los coeficientes de consistencia interna 13

8. Cuándo un coeficiente de fiabilidad es suficientemente alto 15

9. Utilidad de los coeficientes de fiabilidad........ 16

   9.1. Fiabilidad y unidimensionalidad: apoyo a la interpretación *unidimensional* del rasgo medido 16
      9.1.1. Una fiabilidad alta no es prueba inequívoca de que todos los ítems *miden lo mismo.* Necesidad de controles conceptuales........ 18
      9.1.2. Fiabilidad y número de ítems........ 18
      9.1.3. Fiabilidad y simplicidad o complejidad del rasgo medido 19

   9.2. El error típico de la medida........ 19
      9.2.1. Concepto y fórmula del error típico........ 20
      9.2.2. Las puntuaciones *verdaderas*........ 21
      9.2.3. Los *intervalos de confianza de las puntuaciones individuales*........ 22

   9.3. Los coeficientes de correlación corregidos por atenuación........ 22

10. Cuando tenemos un coeficiente de fiabilidad bajo 23
10.1. Inadecuada formulación de los ítems.. 23
10.2. Homogeneidad de la muestra 24
10.3. Definición compleja del rasgo medido 24
10.4. Utilidad del error típico cuando la fiabilidad es baja.. 24

11. La fiabilidad en exámenes y pruebas escolares.. 25
11.1. Fiabilidad y validez... 25
11.2. Fiabilidad y diferencias entre los sujetos.. 25
11.3. Fiabilidad y calificación.. 26

12. Fórmulas de los coeficientes de *consistencia interna* 27
12.1. *Fórmulas basadas en la partición del test en dos mitades*................................ 27
12.1.1. Cómo dividir un test en dos mitades... 27
12.1.2. Fórmulas 27
12.2. Fórmulas de Kuder-Richardson y α de Cronbach 28
12.3. Fórmulas que ponen en relación la fiabilidad y el número de ítems 30
12.3.1. Cuánto aumenta la fiabilidad al aumentar el número de ítems............. 30
12.3.2. En cuánto debemos aumentar el número de ítems para alcanzar una determinada fiabilidad 30
12.4. Estimación de la fiabilidad en una nueva muestra cuya varianza conocemos a partir de la varianza y fiabilidad calculadas en otra muestra.......................... 31
12.5. La fiabilidad en los tests de criterio.. 32

13. Resumen: concepto básico de la fiabilidad en cuanto *consistencia interna* 33

14. Comentarios bibliográficos 34

15. Referencias bibliográficas... 36

1. Conceptos preliminares básicos

Antes de entrar en explicaciones muy precisas y en fórmulas concretas, nos es útil hacer una *aproximación conceptual* a lo que entendemos por fiabilidad en nuestro contexto (los tests, la medición en las ciencias sociales) porque lo que entendemos aquí por fiabilidad es de alguna manera análogo a lo que entendemos por fiabilidad en otras situaciones de la vida corriente. También es útil desde el principio distinguir la *fiabilidad* de conceptos como el de *validez* que utilizamos en los mismos contextos y situaciones y en referencia al uso de los tests.

1. *Equivocidad del concepto de fiabilidad*

El concepto de *fiabilidad,* tal como lo aplicamos en la medición en las ciencias humanas, desemboca en diversas métodos o enfoques de *comprobación* que se traducen en unos *coeficientes de fiabilidad* que a su vez suponen conceptos o definiciones distintas de lo que es la fiabilidad, por lo que tenemos en principio un concepto equívoco más que unívoco. Por esta razón cuando en situaciones aplicadas se habla de la fiabilidad o de coeficientes de fiabilidad, hay que especificar de qué fiabilidad se trata. Esto quedará más claro al hablar de los distintos enfoques, pero conviene tenerlo en cuenta desde el principio.

2. *Fiabilidad y precisión de la medida*

Aun así cabe hablar de un concepto más genérico de fiabilidad con el que se relacionan los otros conceptos más específicos. *En principio* la fiabilidad expresa el grado de ***precisión*** de la medida. Con una fiabilidad alta los sujetos medidos con el mismo instrumento en ocasiones sucesivas hubieran quedado ordenados de manera semejante. Si baja la fiabilidad, *sube el error,* los resultados hubieran variado más de una medición a otra.

Ninguna medición es perfecta; en otro tipo de ámbitos una manera de verificar la precisión es medir lo mismo varias veces, o varios observadores independientes miden lo mismo para obtener una *media* que se estima más precisa que lo que un único observador ha estimado, como cuando se desea comprobar la densidad de una determinada especie animal en un determinado *hábitat*. En la medición psicológica y educacional, que es la que nos interesa, no es posible o no es tan fácil utilizar procedimientos o estrategias que se utilizan más en otros campos de la ciencia; tendremos que buscar otros enfoques para apreciar e incluso cuantificar la precisión de nuestras medidas (como puede ser la precisión de un instrumento para medir conocimientos, actitudes, un rasgo de personalidad, etc.). Lo que importa destacar aquí es la asociación entre los conceptos de *fiabilidad* y *precisión o exactitud*.

3. *Fiabilidad y margen de error en al medida.*

Ya hemos indicado que si fiabilidad significa *precisión*, a menor fiabilidad subirá el *margen de error* de nuestras medidas. En muchas aplicaciones prácticas el interés de los coeficientes de fiabilidad está precisamente en que nos permiten calcular ese margen de error que a su vez nos permiten *relativizar* los resultados individuales, por eso junto a la fiabilidad hay que estudiar *el error típico de la medida* (apartados 9.2 y 11, referido a resultados escolares).

4. *Fiabilidad y predictibilidad*

Otro concepto que nos ayuda a comprender qué entendemos por fiabilidad es el de *consistencia* o *predictibilidad*. Nos *fiamos* de un amigo cuando sabemos *cómo* va a reaccionar ante un problema que le llevemos, y esto lo sabemos porque tenemos experiencias repetidas. De manera análoga un jugador de fútbol es fiable si sabemos de antemano que va a hacer un buen partido, y de nuevo esto lo sabemos porque ya ha jugado bien en otras muchas ocasiones (aunque esto no quiere decir que *siempre* juegue bien).

5. *Fiabilidad y validez*

El concepto de fiabilidad es distinto del concepto de la validez. En el sentido más usual del término (no el único), un instrumento es válido si comprueba o mide aquello que pretendemos medir. Un instrumento puede ser válido, porque mide lo que decimos que mide y queremos medir, pero lo puede *medir* con un *margen de error* grande; con instrumentos parecidos o en mediciones sucesivas hubiéramos obtenido resultados distintos. También puede haber una fiabilidad alta (los sujetos están *clasificados, ordenados,* con poco margen de error) y a la vez el instrumento puede carecer de validez, porque no mide lo que se pretende o lo que se dice que se está midiendo (por ejemplo si un test de *inteligencia* lo que mide realmente es en buena parte *capacidad lectora*, o si un examen supuestamente de *comprensión* lo que se verifica es *memoria y repetición,* etc.).

6. *Fiabilidad y validez: errores sistemáticos y errores aleatorios*

En estos dibujos tenemos dos representaciones gráficas que puede ayudarnos a comprender lo que es *validez* y lo que es *fiabilidad* (figuras 1 y 2).

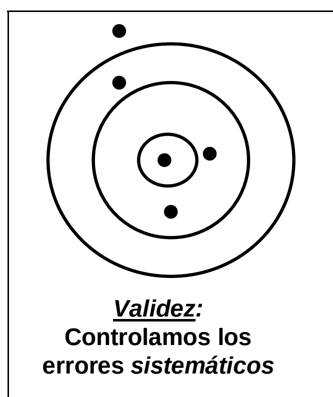


Figura 1

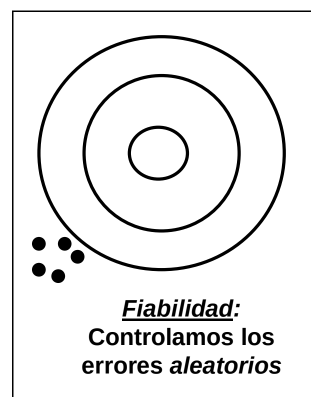


Figura 2

Podemos imaginar a dos tiradores tirando a un blanco. Cuando hay *valide*z (figura 1, *cuando el tiro es válido) se apunta al blanco* aunque puede haber poca precisión. Los errores son *aleatorios* (falta de fiabilidad; fruto de defectos del arma, inestabilidad en el pulso, etc.), pero no son sistemáticos (apuntamos al *blanco que queremos*).

Cuando hay *fiabilidad* hay *precisión* en el tiro. En este ejemplo (figura 2) no hay validez: se apunta *sistemáticamente* fuera del blanco, aunque hay una mayor fiabilidad o precisión (los tiros están más próximos entre sí).

Para evitar los errores aleatorios (y que aumente la fiabilidad) habrá que mejorar el instrumento. Para evitar los errores sistemáticos habrá que *apuntar bien al blanco,* y para eso hay que saber dónde está, y no a otro sitio. La comprobación de la validez sigue otros métodos distintos (análisis del contenido de los ítems, verificar determinadas hipótesis sobre el significado pretendido, etc.) y salvo en casos específicos (como cuando se habla de *validez predictiva)* no se concreta en el cálculo de un coeficiente determinado.

De lo que vamos diciendo se desprende que en primer lugar nos debe preocupar la validez más que la precisión: podemos medir muy bien lo que no queríamos medir (memoria en vez de comprensión, por ejemplo en un examen).

7. *La fiabilidad no es una característica de los instrumentos*

La fiabilidad no es una característica de un instrumento; *es una característica de unos resultados, de unas puntuaciones obtenidas en una muestra determinada.* Esto es importante aunque en el lenguaje habitual nos refiramos a la fiabilidad como si fuera una *propiedad del instrumento;* esta manera de hablar *(este test tiene una fiabilidad de…)* es muy frecuente, pero hay que entender lo que realmente queremos decir. Lo que sucede es que un mismo instrumento puede *medir o clasificar* bien a los sujetos de una muestra, con mucha precisión, y mal, con un margen de error grande, a los sujetos

de otra muestra. Con un mismo instrumento se mide y clasifica mejor cuando los sujetos son muy distintos entre sí, y baja la fiabilidad si la muestra es más homogénea. *La fiabilidad se debe calcular con cada nueva muestra,* sin aducir la fiabilidad obtenida con otras muestras como aval de la fiabilidad del instrumento[1].

Todo esto quedará mejor entendido al examinar las variables que inciden en que un coeficiente de fiabilidad sea alto o bajo, pero es importante ver desde el principio que *en sentido propio* la fiabilidad *no es la propiedad de un determinado instrumento,* sino del conjunto de puntuaciones con él obtenido y que puede variar de una situación a otra.

8. *Fiabilidad y diferencias: teoría clásica de la fiabilidad*

En última instancia lo que nos va a decir un coeficiente de fiabilidad es si el instrumento diferencia adecuadamente a los sujetos en aquello que mide el test o escala. Con un test o escala pretendemos *diferenciar* a los sujetos; establecer quién tiene *más o menos* del rasgo que medimos. Los tests, sobre todo los que miden rasgos psicológicos, no nos serían útiles si de alguna manera no establecieran diferencias entre los sujetos. Ya veremos que, por lo tanto, no hay fiabilidad sin diferencias. Por estas razones la fiabilidad de un test de conocimientos o de un examen (prueba objetiva) no se puede interpretar automáticamente como un indicador de la *calidad* del test, como comentamos en el apartado 7 sobre la interpretación de estos coeficientes cuando se trata de medir conocimientos. A esta teoría de la fiabilidad basada en las diferencias se le suele denominar *teoría clásica de la fiabilidad.*

2. Enfoques y métodos en el cálculo de la fiabilidad

En el cálculo de la fiabilidad hay tres enfoques que, aunque parten de modelos teóricos idénticos o parecidos, siguen procedimientos distintos y sus resultados no pueden interpretarse exactamente del mismo modo; por eso hemos dicho al principio que el concepto de fiabilidad es en cierto modo equívoco. Estos tres enfoques son 1º) el *test-retest,* 2º) el de las *pruebas paralelas* y 3º) los coeficientes de *consistencia interna.*

2.1. Método: Test-retest

Los sujetos responden dos veces al mismo test, dejando entre las dos veces un intervalo de tiempo. ***El coeficiente de correlación entre las dos ocasiones*** es lo que denominamos coeficiente de fiabilidad *test-retest*. El intervalo de tiempo puede ser de días, semanas o meses, pero no tan grande que los sujetos hayan podido cambiar. Una correlación grande indica que en las dos veces los sujetos han quedado *ordenados* de la misma o parecida manera. El intervalo de tiempo debe especificarse siempre (y suele estar en torno a las dos o tres semanas).

a) Este método corresponde al concepto más intuitivo de fiabilidad: un instrumento es fiable si en veces sucesivas aporta los mismos resultados.

b) No tiene sentido utilizarlo cuando está previsto un cambio en los sujetos, o cuando entre la primera y segunda vez se puede dar un aprendizaje. Por esto no es un método apto para comprobar la fiabilidad de un instrumento de medición escolar porque puede haber aprendizaje de una vez a otra, aprendizaje que puede incluso estar provocado por el mismo instrumento. La fiabilidad del tipo test-retest tiene más sentido en la medición de rasgos y actitudes más estables.

c) Este coeficiente de correlación podemos entenderlo como un coeficiente o indicador de ***estabilidad*** o de ***no ambigüedad*** en la medida en que ambas ocasiones los resultados son parecidos

---

[1] El calcular el coeficiente de fiabilidad en cada nueva muestra es una de las recomendaciones de la *American Psychological Association* (Wilkinson and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999); American Psychological Association (2001) y también está recomendado por la política editorial de buenas revistas (Thompson, 1994)

(los sujetos entendieron lo mismo de la misma manera y respondieron de manera idéntica o casi idéntica).

d) Una fiabilidad alta de este tipo no es garantía de una fiabilidad alta con otro de los enfoques, sobre todo con el de *consistencia interna* que veremos enseguida. Puede haber una fiabilidad alta de este tipo, test-retest, con ítems que preguntan cosas muy distintas, con poca *consistencia interna*.

2.2. Método: Pruebas paralelas

Se utiliza cuando se preparan dos versiones del mismo test; los ítems son distintos en cada test pero con ambos se pretende medir lo mismo. En este caso ***el coeficiente de fiabilidad es la correlación entre las dos formas paralelas***, respondidas por los mismos sujetos.

a) Puede interpretarse como un coeficiente o indicador de ***equivalencia*** entre los dos tests: si la correlación es alta, las dos formas del mismo test dan resultados parecidos, ordenan a los sujetos de manera parecida, ambas formas son *intercambiables*. Si la correlación entre las dos formas (respondidas con días u horas de diferencia) es baja, la conclusión más razonable no es que los sujetos han cambiado, sino que las dos formas no están equilibradas en sus contenidos y de alguna manera miden *cosas distintas* o con énfasis distintos.

b) Una confirmación adicional de que las dos formas son realmente *paralelas* es comprobar si la correlación media inter-ítem dentro de cada forma es de magnitud similar, lo mismo que la correlación de los ítems de una forma con los de la otra versión.

c) Este tipo de fiabilidad, o *prueba de equivalencia,* es necesario siempre que se disponga de dos o más versiones del mismo test, y su uso queda en la práctica restringido a esta circunstancia no frecuente.

2.3. Método: Coeficientes de consistencia interna[2]

Este es el enfoque más utilizado y al que le vamos a dar una mayor extensión. Hay que hablar de *enfoque* más que de *método* pues son muchas las posibles fórmulas en que se puede concretar en el cálculo de la fiabilidad. Cuando se habla de fiabilidad sin más matizaciones, hay que entender que se trata de fiabilidad en el sentido de *consistencia interna*.

Lo que expresan directamente estos coeficientes es hasta qué punto las respuestas son lo suficientemente coherentes (relacionadas entre sí) como para poder concluir que *todos los ítems miden lo mismo*, y por lo tanto son *sumables* en una puntuación total única que *representa*, *mide* un rasgo. Por esta razón se denominan coeficientes de *consistencia interna*, y se aducen como garantía de *unidimensionalidad*, es decir, de que un único rasgo subyace a todos los ítems. Hay que advertir sin embargo que un alto coeficiente de fiabilidad *no es prueba* de unidimensionalidad (tratado con más amplitud en el apartado 9.1).

El resto de la teoría sobre la fiabilidad que exponemos a continuación responde fundamentalmente a la fiabilidad entendida como consistencia interna. Cuando se habla de la fiabilidad de un instrumento y no se especifica otra cosa, suele entenderse que se trata de la fiabilidad entendida como *consistencia interna*.

---

[2] Los coeficientes de *consistencia interna* también suelen denominarse coeficientes de *homogeneidad* como si se tratara de términos sinónimos, pero este término (*coeficiente de homogeneidad*) es impropio (como advierte Schmitt, 1996). La *consistencia interna* se refiere a las correlaciones entre los ítems (*relación empírica*) y la homogeneidad se refiere a la *unidimensionalidad* (*relación lógica, conceptual*) de un conjunto de ítems que supuestamente expresan el mismo rasgo.

3. Los coeficientes de *consistencia interna:* concepto y fórmula básica de la fiabilidad

Como punto de partida podemos pensar que cuando observamos diferencias entre los sujetos, estas diferencias, que se manifiestan en que sus puntuaciones totales distintas, se deben

1º En parte a que los sujetos son distintos en aquello que se les está midiendo; si se trata de un examen hay diferencias porque unos saben más y otros saben menos.

2º Las diferencias *observadas* se deben también en parte a lo que llamamos genéricamente *errores de medición;* por ejemplo, en este caso, a preguntas ambiguas, a diferente capacidad lectora de los sujetos, etc.; *no* todo lo que hay de diferencia se debe a que *unos saben más y otros saben menos*.

La *puntuación total* de un sujeto podemos por lo tanto descomponerla así:

$X_t = X_v + X_e$ [1]

X_t = puntuación total de un sujeto, puntuación observada;

X_v = puntuación *verdadera*, que representa lo que un sujeto realmente sabe o siente (depende de qué se esté preguntando o midiendo).

X_e = puntuación debida a *errores de medición*, que puede tener signo *más* o signo *menos.*

Lo que decimos de cada puntuación individual lo podemos decir también de las diferencias entre todos los sujetos:

Diferencias *observadas* entre los sujetos	=	Diferencias *verdaderas;* los sujetos son distintos en lo que estamos midiendo.	+	Diferencias *falsas* *(errores de medición)*

Hablando con propiedad, más que de diferencias concretas hay que hablar de *varianza*, que cuantifica todo lo que hay de diferencia entre los sujetos. La fórmula básica de la fiabilidad parte del hecho de que la varianza se puede descomponer. La varianza de las puntuaciones totales de un test podemos descomponerla así [2]:

$$\sigma^2_t = \sigma^2_v + \sigma^2_e \qquad [2]$$

σ^2_t = *Varianza total,* expresa *todo* lo que hay de diferente en las puntuaciones totales; unos sujetos tienen puntuaciones totales más altas, otros más bajas, etc.; la varianza será mayor si los sujetos difieren mucho entre sí. Si lo que pretendemos con un instrumento de medida es *clasificar, detectar diferencias*, una varianza grande estará asociada en principio a una mayor fiabilidad.

σ^2_v = *Varianza verdadera;* expresa todo lo que hay de diferente debido a que los sujetos son distintos en lo que pretendemos medir, o dicho de otra manera, expresa todo lo que hay de diferente debido a lo que los ítems *tienen en común*, de *relación*, y que es precisamente lo que queremos medir. . El término *verdadero* no hay que entenderlo en un sentido cuasi filosófico, aquí la *varianza verdadera* es la que se debe a *respuestas coherentes* (o respuestas *relacionadas),* y esta coherencia (o *relación verificada)* en las respuestas *suponemos* que se debe a que los ítems *miden lo mismo*

σ^2_e = V*arianza debida a errores de medición,* o debida a que los ítems miden en parte *cosas distintas*, a lo que no tienen en común. Puede haber otras fuentes de error (respuestas descuidadas, falta de motivación al responder, etc.), pero la fuente de error que controlamos es la debida a falta de *relación* entre los ítems, que pueden medir cosas distintas o no muy relacionadas. El *error* aquí viene a ser igual a *incoherencia* en las respuestas, cualquiera que sea su origen (incoherencia sería aquí responder *no* cuando se ha respondido *sí* a un ítem de formulación supuestamente equivalente).

Suponemos que los errores de medición no están relacionados con las puntuaciones verdaderas; no hay *más error* en las puntuaciones más altas o menos en las más bajas y los errores de medición se reparten aleatoriamente; con este supuesto la fórmula [2] es correcta.

La fiabilidad no es otra cosa que la *proporción de varianza verdadera*, y la fórmula básica de la fiabilidad [3] se desprende de la fórmula anterior [2]:

$$r_{11} = \frac{\sigma_v^2}{\sigma_t^2} \qquad [3]$$

Por *varianza verdadera* entendemos lo que acabamos de explicar; la varianza total no ofrece mayor problema, es la que calculamos en los totales de todos los sujetos; cómo hacemos *operativa* la *varianza verdadera* lo veremos al explicar las fórmulas (de Cronbach y Kuder-Richardson). Expresando la fórmula [3] en términos verbales tenemos que

$$\textit{fiabilidad} = \frac{\text{todo lo que discriminan los ítems por lo que tienen de relacionados}}{\text{todo lo que discriminan de hecho al sumarlos en una puntuación total}}$$

o expresado de otra manera

$$\textit{fiabilidad} = \frac{\text{varianza debida a lo que hay de coherente en las respuestas}}{\text{varianza debida tanto a lo que hay de coherente como de no coherente en las respuestas}}$$

Por *respuestas coherentes* hay que entender que no se responde de manera distinta a ítems que supuestamente y según la intención del autor del instrumento, expresan el mismo rasgo a actitud. En una escala de *actitud hacia la música* sería coherente estar de acuerdo con estos dos ítems: *me sirve de descanso escuchar música clásica* y *la educación musical es muy importante en la formación de los niños;* lo coherente es estar de acuerdo con las dos afirmaciones o no estar tan de acuerdo también con las dos. Un sujeto que esté de acuerdo con una y no con la otra es de hecho incoherente según lo que pretende el autor del instrumento (medir la misma actitud a través de los dos ítems). Esta incoherencia *de hecho* no quiere decir que el sujeto no sea coherente con lo que piensa; lo que puede y suele suceder es que los ítems pueden estar mal redactados, pueden ser ambiguos, medir cosas distintas, etc.; por estas razones la fiabilidad hay que verificarla experimentalmente.

En la varianza total (todo lo que hay de diferencias individuales en las puntuaciones totales) influye tanto lo que se responde de manera coherente o relacionada, como lo que hay de incoherente o inconsistente (por la causa que sea); la fiabilidad expresa la proporción de *consistencia* o *coherencia empírica*.

En el denominador tenemos la varianza de los totales, por lo tanto la fiabilidad indica la proporción de varianza debida a lo que los ítems *tienen en común*. Una fiabilidad de .80, por ejemplo, significa que el 80% de la varianza se debe a lo que los ítems tienen en común (o de *relacionado* de hecho).

4. Requisitos para una fiabilidad alta

Si nos fijamos en la fórmula anterior [3] (y quizás con más claridad si nos fijamos en la misma fórmula expresada con palabras), vemos que aumentará la fiabilidad si aumenta el numerador; ahora bien, es importante entender que aumentará el numerador si por parte de los sujetos hay respuestas *distintas y a la vez relacionadas,* de manera que *tendremos una fiabilidad alta:*

1º ***Cuando haya diferencias en las respuestas*** a los ítems, es decir, cuando los ítems *discriminan*; si las respuestas son muy parecidas (todos de acuerdo, o en desacuerdo, etc.) la varianza de los ítems baja y también la fiabilidad;

2º ***Y además los ítems (las respuestas) estén relacionadas entre sí***, hay coherencia, *consistencia interna*; si se responde *muy de acuerdo* a un ítem, se responde de manera parecida a ítems distintos pero que expresan, suponemos, el mismo rasgo; hay una tendencia generalizada responder o en la zona del acuerdo o en la zona del desacuerdo.

Entender cómo estos dos requisitos (respuestas distintas en los sujetos *y* relacionadas) influyen en la fiabilidad es también entender en qué consiste la fiabilidad en cuanto *consistencia interna*. Esto lo podemos ver con facilidad en un ejemplo ficticio y muy simple en el que dos muestras de cuatro sujetos responden a un test de dos ítems con respuestas *sí* o *no* (1 ó 0) (tabla 1).

Grupo A				Grupo B			
	ítem nº 1	ítem nº 2	Total		ítem nº 1	ítem nº 2	Total
sujeto a	1	1	2	sujeto a	0	1	1
sujeto b	1	1	2	sujeto b	1	0	1
sujeto c	0	0	0	sujeto c	1	1	2
sujeto d	0	0	0	sujeto d	0	0	0
$\bar{X}$ =	.50	.50	1	$\bar{X}$ =	.50	.50	1
σ =	.50	.50	1	σ =	.50	.50	.70
σ^2 =	.25	.25	1	σ^2 =	.25	.25	.50
Correlación entre los ítems:		r =	1	Correlación entre los ítems:		r =	0
Fiabilidad (Kuder-Richardson 20)		r_{11} =	1	Fiabilidad (Kuder-Richardson 20)		r_{11} =	0
(máxima fiabilidad o consistencia interna posible)				(nula consistencia interna)			

Tabla 1

Podemos pensar que se trata de una escala de *integración familiar* compuesta por dos ítems y respondida por dos grupos de cuatro sujetos cada uno. Los ítems en este ejemplo podrían ser:

1. *En casa me lo paso muy bien con mis padres* [sí=1 y no =0]
2. *A veces me gustaría marcharme de casa* [sí = 0 y no = 1]

En estos ejemplos podemos observar:

1º Las desviaciones típicas (lo mismo que las varianzas, σ^2) de los ítems son idénticas en los dos casos, además son las máximas posibles (porque el 50% está de acuerdo y el otro 50% está en desacuerdo, máxima dispersión). Desviaciones típicas grandes en los ítems (lo que supone que distintos sujetos responden de distinta manera al mismo ítem) contribuyen a aumentar la fiabilidad, pero vemos que no es condición suficiente: con las mismas desviaciones típicas en los ítems el coeficiente de fiabilidad es 1 (grupo A) en un caso y 0 en otro (grupo B).

2º La diferencia entre los grupos A y B está en las correlaciones inter-ítem: la máxima posible en A (r = 1), y la más baja posible en B (r = 0). La correlación es grande cuando las respuestas son *coherentes*, cuando se responde básicamente de la misma manera a todos los ítems; la correlación es pequeña cuando las respuestas son *incoherentes*.

Cuando las respuestas son coherentes (simplificando: unos dicen que *sí* a todo y otros dicen que *no* a todo), la puntuación total está más diversificada porque se acumulan puntuaciones muy altas o muy bajas en los ítems; consecuentemente la desviación típica (o la varianza) de los totales será mayor. Con respuestas *diferentes* y además *coherentes*, los sujetos quedan *más diversificados*, mejor *clasificados* por sus puntuaciones totales, y esta diversidad de los totales se refleja en una mayor desviación típica o varianza.

Esta *diversidad coherente* de las respuestas (y que la vemos de manera exagerada en el grupo A del ejemplo anterior) queda recogida en la fórmula de la fiabilidad o de *consistencia interna*.

Para que suba la fiabilidad hace falta por lo tanto lo que ya hemos indicado antes:

1º que unos y otros sujetos respondan de manera *distinta* a los ítems
2º y que *además* esas respuestas a los ítems sean coherentes.

Si esto es así, las diferencias en los totales se deberán a que los sujetos han respondido de manera *distinta y coherente* a los distintos ítems. Esto hace que los totales sean distintos, para unos sujetos y otros, según *tengan más o menos* del *rasgo* que deseamos *medir*. unos van acumulando valores altos en sus respuestas, y otros van acumulando valores bajos.

Lo que significa la fiabilidad, y las condiciones de una fiabilidad alta, podemos verlo en otro ejemplo ficticio (tabla 2). Imaginemos que se trata ahora de una escala (de *actitud hacia la música)* de tres ítems, con respuestas continuas de 1 a 5 (de *máximo desacuerdo* a *máximo acuerdo)* respondida por seis sujetos:

	ítems			
sujetos:	1. *La música es lo que más me gusta*	2. *Me encanta escuchar música*	3. *Me entusiasma la ópera*	Total
1. Un melómano	5	5	5	15
2. Otro melómano	5	5	5	15
3. Un aficionado	3	4	4	11
4. Otro aficionado, le aburre la ópera	4	5	2	11
5. No le gusta la música	1	3	2	6
6. Le aburre la música	1	2	1	4
Varianzas:	2.805	1.33	2.472	17.22

Tabla 2

Qué vemos en estos datos:

1. Los ítems *miden lo mismo conceptualmente;* al menos es lo que intentamos al redactarlos;

2. *Los sujetos son distintos* en las respuestas a *cada ítem*, por eso hay *varianza* (diferencias) en los ítems; a unos les gusta más la música, a otros menos;

3. Los ítems están *relacionados:* si tomamos los ítems de dos en dos vemos que los sujetos tienden a puntuar alto en todos o bajo en todos (más o menos); esta relación podemos verificarla experimentalmente calculando los coeficientes de correlación: r_{12} = .95, r_{13} = .81 y r_{23} = .734 (en ejemplos *reales*, con más ítems y más sujetos, no suelen ser tan altos).

4. Consecuentemente el puntuar alto en un ítem supone un total más alto en toda la escala; esto podemos verificarlo experimentalmente calculando la correlación de cada ítem con la suma de los otros dos *(correlación ítem-total):* r_{1t} = .93, r_{2t} = .88 y r_{3t} = .79. Un procedimiento que nos da la misma información es comparar en cada ítem a los sujetos con totales más altos y totales más bajos; si los mismos ítems diferencian simultáneamente a los mismos sujetos, es que los ítems están relacionados.

5. Los sujetos van acumulando puntuaciones altas o bajas en cada ítem, por lo tanto quedan muy diferenciados en la puntuación total: están bien *ordenados* o *clasificados.*

6. Nos encontramos con una *coherencia global* en las respuestas, *todos los ítems están relacionados;* esta coherencia global es la que estimamos en los coeficientes de fiabilidad (de *consistencia interna;* en este caso el coeficiente de fiabilidad es α = .924.

7. Esta relación entre los ítems es la que comprobamos experimentalmente y nos permite sumarlos en una sola puntuación total porque nos confirma (aunque no necesariamente) que todos *miden lo mismo*. Si un ítem no está claramente relacionado con los demás, puede ser que esté *midiendo otra cosa.*

8. La relación *conceptual (homogeneidad de los ítems)* la suponemos (procuramos que todos los ítems expresen el mismo rasgo, aunque podemos equivocarnos), pero la comprobamos *empíricamente* en cada ítem (mediante la correlación de cada ítem con todos los demás) y en el conjunto de todo el instrumento (coeficiente de fiabilidad).

Incidentalmente hacemos una observación: si en vez del ítem *me entusiasma la ópera* (que claramente expresa gusto por la música) ponemos *en mi casa tengo un piano,* que *podría* expresar gusto por la música pero también, y con más claridad, indica nivel económico (algo *distinto* al gusto por la música, con unas respuestas *no sumables* con las demás), y los dos melómanos del ejemplo son además ricos y tienen un piano en casa y los dos a quienes no gusta o gusta menos la música son de

nivel económico inferior y por supuesto no tienen un piano en su casa, tendríamos que este ítem, *en mi casa tengo un piano,* está contribuyendo a la fiabilidad de la escala sin que podamos decir que está midiendo *lo mismo* que los demás. *Los números no entienden de significados,* de ahí la insistencia en los controles conceptuales.

9. El coeficiente de fiabilidad aumenta por lo tanto:

a) si hay *diferencias en las respuestas a cada ítem*
b) y si *además hay relación entre los ítems* (es decir, hay coherencia en las respuestas).

10. La fiabilidad supone también que los sujetos son *distintos en aquello que es común a todos los ítems.* El mismo test o escala, con los *mismos* ítems, puede tener una fiabilidad alta en una muestra y baja en otra: si todos responden a los ítems de idéntica manera: a) los ítems tendrán varianzas pequeñas y b) interrelaciones pequeñas, y por lo tanto bajará la fiabilidad. La fiabilidad viene a expresar la capacidad del instrumento para discriminar, para diferenciar a los sujetos a través de sus respuestas a todos los ítems. Es más probable encontrar una fiabilidad alta en una muestra grande, porque es más probable también que haya sujetos más extremos en lo que estamos midiendo. En sentido propio la fiabilidad no es una propiedad del test o escala, sino de las puntuaciones obtenidas con el instrumento en una muestra dada.

5. Las fórmulas de Kuder Richardson 20 y α de Cronbach

Las dos fórmulas posiblemente más utilizadas son las de Kuder-Richardson 20 y el coeficiente α de Cronbach. En realidad se trata de la misma fórmula, una (Kuder-Richardson) expresada para ítems dicotómicos (*unos* y *ceros*) y otra (Cronbach) para ítems continuos. Los nombres distintos se deben a que los autores difieren en sus modelos teóricos, aunque estén relacionados, y los desarrollaron en tiempos distintos (Kuder y Richardson en 1937, Cronbach en 1951).

Para hacer operativa la fórmula [3] $\quad r_{11} = \dfrac{\sigma_v^2}{\sigma_t^2}$

El *denominador* no ofrece mayor problema, se trata de la varianza de las puntuaciones totales del test o instrumento utilizado.

El *numerador,* o *varianza verdadera,* lo expresamos a través de la *suma de las covarianzas de los ítems.* Es útil recordar aquí qué es la *co-varianza*. Conceptualmente la *co-varianza* es lo mismo que la *co-relación;* en el coeficiente de correlación utilizamos puntuaciones típicas y en la covarianza utilizamos puntuaciones directas, pero en ambos casos se expresa lo mismo y si entendemos qué es la correlación, entendemos también qué es la covarianza o *variación conjunta.* La *varianza verdadera* la definimos operativamente como la *suma de las covarianzas de los ítems*.

La covarianza entre dos ítems expresa lo que dos ítems *discriminan por estar relacionados,* esto es lo que denominamos en estas fórmulas *varianza verdadera*, por lo tanto la fórmula [3] podemos expresarla poniendo en el numerador la *suma de las covarianzas entre los ítems:*

$$r_{11} = \frac{\Sigma\sigma_{xy}}{\sigma_t^2} \quad [4] \qquad \text{o lo que es lo mismo} \qquad r_{11} = \frac{\Sigma r_{xy}\sigma_x\sigma_y}{\sigma_t^2} \quad [5] \qquad \text{ya que } \sigma_{xy} = r_{xy}\sigma_x\sigma_y$$

La covarianza entre dos ítems (σ_{xy}) es igual al producto de su correlación (r_{xy}) por sus desviaciones típicas (σ_x y σ_y): ahí tenemos la *varianza verdadera: diferencias* en las respuestas a los ítems (expresadas por las desviaciones típicas) y *relacionadas* (relación expresada por los coeficientes de correlación entre los ítems). Se trata por lo tanto de *relaciones empíricas*, verificadas, no meramente lógicas o conceptuales.

Esta fórmula [5] de la fiabilidad no es, por supuesto cómoda para calcularla (tenemos otras alternativas) pero pone de manifiesto qué es lo que influye en la fiabilidad, por eso es importante.

Aumentará la fiabilidad si aumenta el numerador. Y lo que tenemos en el numerador (fórmula 5) es la *suma de las covarianzas de los ítems* ($\Sigma\sigma_{xy} = \Sigma r_{xy}\sigma_x\sigma_y$) que expresa a) *todo lo que discriminan los ítems* (y ahí están sus desviaciones típicas) y b) *por estar relacionados* (y tenemos también las correlaciones inter-ítem).

Si nos fijamos en la fórmula [5] vemos que si los ítems no discriminan (no establecen diferencias) sus desviaciones típicas serán pequeñas, bajará el numerador y bajará la fiabilidad.

Pero no basta con que haya diferencias en los ítems, además tienen que estar relacionados; la correlación entre los ítems también está en el numerador de la fórmula [5]: si las desviaciones son grandes (como en el grupo B de la tabla 1) pero los ítems no están relacionados (= respuestas no coherentes), bajará la fiabilidad, porque esa *no relación* entre los ítems hace que las puntuaciones totales estén menos diferenciadas, como sucede en el grupo B. En este caso vemos que cuando las desviaciones de los ítems son muy grandes, pero la correlación inter-ítem es igual a 0, la fiabilidad es también igual a 0.

La fiabilidad expresa por lo tanto *cuánto hay de diferencias en los totales debidas a respuestas coherentes* (o proporción de *varianza verdadera* o debida a que los ítems *están relacionados).* Por eso se denomina a estos coeficientes *coeficientes de consistencia interna*: son mayores cuando las relaciones entre los ítems son mayores. La expresión *varianza verdadera* puede ser equívoca; en este contexto varianza verdadera es la debida a que los ítems *están relacionados*, son respondidos de manera básicamente coherente, pero no prueba o implica que *de verdad* todos los ítems midan lo mismo.

Esta *relación empírica*, verificable, entre los ítems nos sirve para *apoyar* o confirmar (pero no *probar*) la *relación conceptual* que debe haber entre los ítems (ya que pretendidamente *miden lo mismo)*, aunque esta *prueba* no es absoluta y definitiva y requerirá matizaciones adicionales (dos ítems pueden estar muy relacionados entre sí sin que se pueda decir que *miden lo mismo*, como podrían ser edad y altura).

La fórmula [4] puede transformarse en otra de cálculo más sencillo. Se puede demostrar fácilmente que *la varianza de un compuesto* (como la varianza de los totales de un test, que está compuesto de una serie de ítems que se suman en una puntuación final) *es igual a la suma de las covarianzas entre los ítems* (entre las partes del compuesto) *más la suma de las varianzas de los ítems:*

$$\sigma_t^2 = \Sigma\sigma_{xy} + \Sigma\sigma_i^2 \qquad [6] \qquad \text{de donde } \Sigma\sigma_{xy} = \sigma_t^2 - \Sigma\sigma_i^2$$

y sustituyendo en [4] tenemos que

$$r_{11} = \frac{\sigma_t^2 - \Sigma\sigma_i^2}{\sigma_t^2} = \frac{\sigma_t^2}{\sigma_t^2} - \frac{\Sigma\sigma_i^2}{\sigma_t^2} \qquad \text{de donde} \qquad r_{11} = 1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2} \qquad [7]$$

La fórmula que sin embargo utilizamos es esta otra y que corresponde al coeficiente α de Cronbach:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2}\right) \qquad [8]$$

k = número de ítems
$\Sigma\sigma_i^2$ = suma de las varianzas de los ítems
σ_t^2 = varianza de los totales.

La expresión [k/(k-1)] (k= número de ítems) la añadimos para que el valor máximo de este coeficiente pueda llegar a la unidad. El segundo miembro de esta fórmula, que es el que realmente cuantifica la proporción de varianza debida a lo que los ítems tienen en común o de relacionado, puede alcanzar un valor máximo de [(k-1)/k] y esto solamente en el caso improbable de que todas las varianzas y covarianzas sean iguales. Como [(k-1)/k] x [k/(k-1)]= 1, al añadir a la fórmula el factor [k/(k-1)] hacemos que el valor máximo posible sea 1.

La fórmula [8], tal como está expresada, corresponde al α de Cronbach (para ítems continuos); en la fórmula Kuder-Richardson 20 (para ítems dicotómicos, respuesta 1 ó 0) sustituimos $\Sigma\sigma_i^2$ por Σpq pues pq es la varianza de los ítems dicotómicos (p = proporción de *unos* y q = proporción de *ceros*).

La parte de la fórmula [8] que realmente clarifica el sentido de la fiabilidad está en el segundo miembro que, como hemos visto, equivale a $\Sigma r_{xy}\sigma_x\sigma_y/\sigma_t^2$ (suma de las covarianzas de todos los ítems dividida por la varianza de los totales, fórmulas 4 y 5).

6. Factores que inciden en la *magnitud* del coeficiente de fiabilidad

Es útil tener a la vista los factores o variables que inciden en coeficientes de fiabilidad altos. Cuando construimos y *probamos* un instrumento de medición psicológica o educacional nos interesa que su fiabilidad no sea baja y conviene tener a la vista qué podemos hacer para obtener coeficientes altos. Además el tener en cuenta estos factores que inciden en la magnitud del coeficiente de fiabilidad nos ayuda a interpretar casos concretos.

En general los coeficientes de fiabilidad tienden a aumentar:

1º ***Cuando la muestra es heterogénea;*** es más fácil clasificar a los sujetos cuando son muy distintos entre sí. Con muestras de sujetos muy parecidos en el rasgo que queremos medir, todos responderán de manera parecida, y las varianzas de los ítems y sus intercorrelaciones serán pequeñas.

2º. ***Cuando la muestra es grande*** porque en muestras grandes es más probable que haya sujetos muy distintos (es la heterogeneidad de la muestra, y no el número de sujetos, lo que incide directamente en la fiabilidad); aunque también podemos obtener un coeficiente alto en muestras pequeñas si los sujetos son muy diferentes en aquello que es común a todos los ítems y que pretendemos medir.

3º ***Cuando las respuestas a los ítems son más de dos*** (mayor probabilidad de que las respuestas difieran más, de que se manifiesten las diferencias que de hecho existen). Cuando el número de respuestas supera la capacidad de discriminación de los sujetos, la fiabilidad baja porque las respuestas son más inconsistentes; en torno a 6 ó 7, e incluso menos, suele situarse el número óptimo de respuestas. Lo más claro experimentalmente es que la fiabilidad sube al pasar de dos respuestas a tres.

4º ***Cuando los ítems son muchos*** (más oportunidad de que los sujetos queden más diferenciados en la puntuación total) aunque un número de ítems grande puede dar una idea equívoca de la homogeneidad del instrumento como indicaremos más adelante (*muchos* ítems *poco* relacionados entre sí pueden tener una fiabilidad alta sin que quede muy claro qué se está midiendo).

5º ***Cuando la formulación de los ítems es muy semejante***, muy repetitiva (si hay diferencias entre los sujetos, aparecerán en todos los ítems y subirán sus intercorrelaciones) aunque ésta no es una característica necesariamente deseable en un instrumento (que mediría un constructo definido con límites muy estrechos). En general los constructos o rasgos definidos con un nivel alto de complejidad requerirán ítems más diversificados y la fiabilidad tenderá a ser menor.

7. Interpretación de los coeficientes de consistencia interna

Basándonos en estas fórmulas y en sus modelos teóricos, estos coeficientes podemos interpretarlos de las siguientes maneras (unas interpretaciones se derivan de las otras):

1***. Expresa*** directamente lo que ya hemos indicado: ***la proporción de varianza debida a lo que los ítems tienen de relacionado,*** de común; un coeficiente de .70 indica que el 70% de la varianza (diferencias en los totales, que es lo que cuantifica la varianza) se debe a lo que los ítems tienen *en común* (de estar relacionado, de coherencia en las respuestas), y un 30% de la varianza se debe a

errores de medición o a lo que *de hecho* tienen los ítems de *no relacionado*. De esta interpretación podemos decir que es una *interpretación literal,* que se desprende directamente de la lectura de la fórmula *(Suma de covarianzas/Varianza total).*

Estos coeficientes, dicho en otras palabras, expresan en qué grado los ítems discriminan o diferencian a los sujetos *simultáneamente*. De alguna manera *son un indicador de relación global entre los ítems* (aunque no equivalen a la correlación media entre los ítems).

2. Consecuentemente interpretamos estos coeficientes como índices o ***indicadores de la homogeneidad de los ítems*** (es decir, de que *todos los ítems miden lo mismo,* por eso se denominan coeficientes de *consistencia interna);* pero esto es ya una interpretación: suponemos que si las respuestas están relacionadas es porque los ítems *expresan o son indicadores del mismo rasgo,* ***aunque no hay que confundir relación empírica*** (verificada, relación de hecho en las respuestas y es esto lo que cuantificamos con estas fórmulas) ***con homogeneidad conceptual***. Esta relación o consistencia interna comprobada de los ítems es la que *legitima* su suma en una puntuación total, que es la que utilizamos e interpretamos como *descriptor del rasgo* (ciencia, una actitud, un rasgo de personalidad, etc.) que suponemos presente en todos los ítems.

3. ***Son una estimación del coeficiente de correlación que podemos esperar con un test similar***, con el mismo número y tipo de ítems.

Esta interpretación se deriva directamente del modelo teórico propuesto por Cronbach. De un *universo* o población de posibles ítems hemos escogido una *muestra de ítems* que es la que conforma nuestro instrumento. Si la fiabilidad es alta, con *otra muestra de ítems* de *la misma población de ítems* obtendríamos unos resultados semejantes (los sujetos quedarían ordenados de manera similar).

Un uso importante de estos coeficientes es poder comunicar hasta qué punto los resultados obtenidos con un determinado instrumento son *repetibles,* si con un test semejante los resultados serían similares. Es en este sentido es un indicador de la *eficacia* del instrumento. Si estos coeficientes son una estimación de la correlación del test con otro similar, podemos concluir que con *otro test* semejante los sujetos hubieran quedado *ordenados*, clasificados, de la misma manera.

4. En términos generales el coeficiente de fiabilidad ***nos dice si un test discrimina adecuadamente***, ***si clasifica bien a los sujetos***, si detecta bien las diferencias que existen entre los sujetos de una muestra. Diferencias ¿En qué? En aquello que es *común a todos los ítems* y que es lo que pretendemos medir. Es más, sin diferencias entre los sujetos no puede haber un coeficiente de fiabilidad alto. La fiabilidad es una característica positiva siempre que interese *detectar diferencias* que suponemos que existen. Esto sucede cuando medimos rasgos de personalidad, actitudes, etc., *medir* es, de alguna manera, *establecer diferencias*.

5. Una ***observación importante***: la interpretación de estos coeficientes, como característica *positiva* o *deseable,* puede ser distinta cuando se trata de comprobar *resultados escolares* en los que no hay diferencias o no se pretende que existan, por ejemplo en un examen de *objetivos mínimos,* o si se trata de verificar si *todos* los alumnos han conseguido determinados objetivos. A la valoración de la fiabilidad en exámenes y pruebas escolares le dedicamos más adelante un comentario específico (apartado 11).

La valoración de una fiabilidad alta como característica positiva o de *calidad* de un test es más clara en los tests de *personalidad, inteligencia,* etc., o en las *escalas de actitudes:* en estos casos pretendemos *diferenciar* a los sujetos, captar las diferencias que de hecho se dan en cualquier rasgo; digamos que en estos casos las diferencias son esperadas y legítimas. Además en este tipo de tests también pretendemos *medir* (en un sentido analógico) *un único rasgo* expresado por todos los ítems, mientras que en el caso de un *examen* de conocimientos puede haber habilidades muy distintas, con poca relación entre sí, en el mismo examen (aunque tampoco esto es lo más habitual).

Aun con estas observaciones, en un examen largo, tipo test, con muchos o bastantes alumnos, entre los que esperamos legítimamente que haya diferencias, una fiabilidad baja sí puede ser un indicador de baja calidad del instrumento, que no recoge diferencias que probablemente sí existen.

6. ***Índice de precisión.*** Hemos visto que el coeficiente de fiabilidad expresa una *proporción,* la proporción de *varianza verdadera* o varianza debida a lo que los ítems tienen en común. También sabemos que un coeficiente de correlación elevado al cuadrado (r^2, *índice de determinación)* expresa una proporción (la proporción de varianza compartida por dos variables). Esto quiere decir que la raíz cuadrada de una proporción equivale a un coeficiente de correlación (si r^2 = proporción, tenemos que $\sqrt{\textit{proporción}} = r$).

En este caso la raíz cuadrada de un coeficiente de fiabilidad equivale al coeficiente de correlación entre las puntuaciones *obtenidas* (con nuestro instrumento) y las puntuaciones *verdaderas* (obtenidas con un *test ideal* que midiera lo mismo). Este coeficiente se denomina ***índice de precisión*** (también índice, no coeficiente, de fiabilidad).

índice de precisión ó $r_{\text{observadas.verdaderas}} = \sqrt{r_{11}}$ [9]

Una fiabilidad de .75 indicaría una correlación de .86 $(=\sqrt{.75})$ con las *puntuaciones verdaderas*. Este índice expresa el valor máximo que puede alcanzar el coeficiente de fiabilidad. No es de mucha utilidad, pero se puede utilizar junto con el coeficiente de fiabilidad.

7. La interpretación del coeficiente de fiabilidad se complementa con el cálculo y uso del ***error típico o margen de error****;* es la *oscilación probable* de las puntuaciones si los sujetos hubieran respondido a una serie de *tests paralelos*; a mayor fiabilidad (a mayor *precisión)* bajará la magnitud del error probable. Tratamos del *error típico* en el apartado siguiente; el error típico, como veremos, puede ser de utilidad más práctica que el coeficiente de fiabilidad.

8. Cuándo un coeficiente de fiabilidad es suficientemente alto

Esta pregunta no tiene una respuesta nítida; cada coeficiente hay que valorarlo en su situación: tipo de instrumento (define un rasgo muy simple o muy complejo), de muestra (muy homogénea o más heterogénea) y uso pretendido del instrumento (mera investigación sobre grupos, o toma de decisiones sobre sujetos).

En la práctica la valoración depende sobre todo del uso que se vaya a hacer del instrumento (de las puntuaciones con él obtenidas). Como *orientación* podemos especificar tres usos posibles de los tests y algunos valores orientadores (tabla 3).

	toma de decisiones sobre individuos	descripción de grupos *feedback* a un grupo	investigación teórica; investigación en general
a) .85 o mayor..............................	sí	sí	sí
b) entre .60 y .85..........................	cuestionable	sí	sí
c) inferior a .60	no	cuestionable	sí, cuestionable…

Tabla 3

Estas valoraciones, como otras similares que pueden encontrarse en libros de texto y en diversos autores, son sólo orientadoras[3]. Lo que se quiere poner de manifiesto es que no es lo mismo *investigar*

---

[3] Nunnally (1978) propone un mínimum de .70; para Guilford (1954:388-389) una fiabilidad de sólo .50 es suficiente para investigaciones de carácter básico; Pfeiffer, Heslin y Jones (1976) y otros indican .85 si se van a tomar decisiones sobre sujetos concretos; en algunos tests bien conocidos (de Cattell) se citan coeficientes inferiores a .50 (Gómez Fernández,

(comparar medias de grupos, etc.) que tomar decisiones sobre individuos. Si se van a *tomar decisiones sobre sujetos concretos* (como aprobar, excluir, recomendar tratamiento psiquiátrico, etc.) hay que proceder con más *cautela,* teniendo en cuenta además que no todas las posibles decisiones son de igual importancia. Cuando baja la fiabilidad sube el *error típico* (o margen de error en la puntuación individual) que con una forma paralela del mismo test o en otra situación, etc., podría ser distinta. *Los grupos son más estables que los individuos,* y el margen de error que pueda haber es de menor importancia (el *error típico de la media* es menor que la desviación típica de la muestra).

Por lo demás si se trata de tomar decisiones sobre individuos concretos se puede tener en cuenta el *error típico* y tomar la decisión en función de la *banda de posibles puntuaciones individuales* más que en función de la puntuación concreta obtenida de hecho; de esta manera asumimos la menor fiabilidad del instrumento. Por otra parte tampoco se suelen tomar decisiones importantes en función del resultado de un único test.

En el caso de informar sobre grupos se pueden especificar los intervalos de confianza de la media (margen de error o de oscilación de la media, que se verá en el lugar apropiado).

Los valores del coeficiente de fiabilidad oscilan entre 0 y 1, pero ocasionalmente podemos encontrar valores negativos, simplemente porque no se cumplen en un grado apreciable las condiciones de estos modelos (Black, 1999:286); en este caso (valor negativo) podemos interpretar este coeficiente como *cero*[4].

9. Utilidad de los coeficientes de fiabilidad

Vamos a fijarnos en tres ventajas o usos frecuentes de estos coeficientes:

1º Nos confirman *en principio* que *todos los ítems miden lo mismo,* y de hecho estos coeficientes se utilizan como un *control de calidad,* aunque esta interpretación es discutible y habrá que entenderla y relativizarla. Más bien habría que decir que un coeficiente alto de fiabilidad apoya (pero no prueba) la hipótesis de que todos los ítems miden básicamente el mismo rasgo o atributo.

2º Los coeficientes de fiabilidad permiten calcular el *error típico de las puntuaciones individuales;* este *error típico* puede incluso ser de un interés mayor que el coeficiente de fiabilidad porque tiene aplicaciones prácticas como veremos en su lugar.

3º Los coeficientes de fiabilidad obtenidos nos permiten estimar los coeficientes de correlación que hubiéramos obtenido entre dos variables si su fiabilidad fuera perfecta (y que se denominan coeficientes de correlación *corregidos por atenuación).*

9.1. Fiabilidad y unidimensionalidad: apoyo a la interpretación *unidimensional* del rasgo medido

Como vamos exponiendo, la *consistencia interna* que manifiesta el coeficiente de fiabilidad *apoya* (pero no *prueba)* la interpretación de que todos los ítems *miden lo mismo* (es lo que entendemos

1981). No hay un valor mínimo *sagrado* para aceptar un coeficiente de fiabilidad como adecuado; medidas con una fiabilidad relativamente baja pueden ser muy útiles (Schmitt, 1996). Por otra parte coeficientes muy altos; pueden indicar excesiva *redundancia* en los ítems (muy repetitivos) por esta razón hay autores que recomiendan un máximum de .90 (Streiner, 2003). Como referencia adicional podemos indicar que la *fiabilidad media* en artículos de buenas revistas de Psicología de la Educación está en torno a .83 (Osborne, 2003).

4 Valores negativos del coeficiente de fiabilidad pueden encontrarse cuando hay substanciales correlaciones negativas entre los ítems; esto puede suceder cuando está mal la clave de corrección y hay ítems con una formulación positiva y negativa que tienen la misma clave; también puede suceder que los ítems realmente miden constructos distintos y no hay suficiente *varianza compartida;* en estos casos la fiabilidad puede considerarse igual a cero (Streiner, 2003).

por *unidimensionalidad;* que el instrumento mide un único rasgo bien definido); esto es lo que en principio se pretende cuando se construye un test o escala.

Ésta es la interpretación y valoración más común de estos coeficientes. Simplificando, lo que decimos es esto: si unos sujetos tienden a estar de acuerdo con todos los ítems y otros responden en la zona del desacuerdo a los mismos ítems, esta coherencia de las respuestas nos dice que todos los ítems *miden el mismo rasgo*. Esta interpretación, que es válida en principio, hay que *relativizarla,* porque en la fiabilidad influyen variables ajenas a la redacción de los ítems, que por otra parte pueden ser *buenos* (con criterios conceptuales) pero no para cualquier muestra o para cualquier finalidad.

El interpretar una fiabilidad alta como indicador claro de que todos los ítems *miden lo mismo* no se puede aceptar ingenuamente; *el coeficiente de fiabilidad no es una medida de unidimensionalidad.* Esto es importante porque precisamente se aduce este coeficiente como prueba de que los ítems miden lo mismo, de que todos los ítems expresan bien un mismo rasgo, y esto no está siempre tan claro.

Por otra parte (como ya se ha indicado en el nº 7) una de las interpretaciones *standard* de estos coeficientes (en la misma *línea* de apoyo a la *unidimensionalidad* del test) es que expresan la correlación que obtendríamos con un *test paralelo.* Podemos concebir un test (o escala de actitudes, etc.) como compuesto por una *muestra aleatoria* de ítems tomada de un universo o población de ítems que miden lo mismo: la fiabilidad indicaría la correlación de nuestro test con otro de idéntico número de ítems tomados del mismo universo.

En primer lugar no hay un valor óptimo del coeficiente de fiabilidad y por otra parte esta interpretación (derivada del *modelo* de Cronbach) supone al menos una condición que no suele darse en la práctica: que todas las correlaciones *ítem-total* son de la misma magnitud. En la práctica es preferible hablar de una *estimación* de esa correlación, que será más exacta si somos muy restrictivos en la selección de los ítems.

Hay que matizar la interpretación de estos coeficientes porque no dependen exclusivamente de la *redacción de los ítems,* de la *complejidad* o *simplicidad* de la definición del rasgo que queremos medir, y además (y frecuentemente *sobre todo)* influyen en la fiabilidad *características de la muestra*. Hablando con propiedad, la fiabilidad ya sabemos que no es una característica del instrumento de medición sino de las puntuaciones con él obtenidas en una situación dada y con una muestra determinada.

En estas observaciones nos fijamos sobre todo en los coeficientes de fiabilidad más bien *altos,* porque no indican necesariamente que el instrumento es *bueno,* también prestaremos atención a los coeficientes *bajos,* que pueden tener su explicación e incluso ser compatibles con un *buen instrumento*.

Vamos a explicar por qué un *coeficiente alto* no expresa necesariamente que los ítems son suficientemente homogéneos como para concluir que *todos miden lo mismo,* que hay suficiente *homogeneidad conceptual* como para sumarlos en una única puntuación que refleja lo un sujeto tiene del *rasgo* que estamos midiendo y que consideramos expresado por la formulación de los ítems.

Nos fijaremos en tres puntos:

1º) Esta *consistencia interna* que cuantifican los coeficientes de fiabilidad expresa una *relación de hecho, estadística, empírica,* entre los ítems, pero la relación empírica no supone *necesariamente* que hay *coherencia conceptual* (que todos expresan bien el mismo rasgo).

2º) Una fiabilidad alta puede deberse a un número grande de ítems que en ocasiones no se prestan a una interpretación clara como descriptores de un *único* rasgo, bien definido.

3º) Una fiabilidad alta puede deberse también a una concepción del rasgo muy limitada, expresada a través de ítems de contenido casi idéntico, muy repetitivos.

Todo esto hay que tenerlo en cuenta para *valorar* estos coeficientes y no dar necesariamente por *bueno* un instrumento porque hemos obtenido una *fiabilidad alta*[5].

9.1.1. Una fiabilidad alta no es prueba inequívoca de que todos los ítems *miden lo mismo:* necesidad de controles conceptuales

Puede suceder que los ítems estén relacionados pero que expresen *cosas distintas* (o *suficientemente distintas)* y que sea cuestionable el sumarlos como si realmente *midieran lo mismo*; al menos esa puntuación total puede no ser de interpretación clara.

En una escala de *actitud hacia la música* podemos pensar en estos dos ítems:	1. *En mi tiempo libre me gusta escuchar música* 2. *En mi casa tenemos un piano*.

Estos dos ítems son ejemplo exagerado (porque obviamente no describen el mismo rasgo), pero es claro para ilustrar que relación empírica (la que expresan estos coeficientes de fiabilidad) no es lo mismo que relación conceptual (que de entrada todos los ítems midan un rasgo interpretable).

Si a los que más les gusta la música tienen además un piano en casa, obtendremos una correlación alta entre estos dos ítems, y si además las varianzas de los ítems son grandes, obtendremos una fiabilidad alta (como en el grupo A del ejemplo puesto antes en la tabla 1: tienen un piano y además les gusta la música, o no tienen piano y además no les gusta mucho escuchar música). En este caso sería discutible considerar los dos ítems *homogéneos* como si *midieran lo mismo*, a pesar de un coeficiente de fiabilidad alto. El *tener un piano en casa* mide o expresa nivel económico aunque el tener un piano en casa coincida *de hecho* (no necesariamente pero tendría su lógica) con una actitud favorable hacia la música. Hace falta un control *cualitativo* y no meramente estadístico de la homogeneidad de los ítems.

Además de la fiabilidad, tenemos que considerar la *homogeneidad conceptual*. Aunque esta *homogeneidad conceptual* la suponemos (al menos es lo que se intenta), un índice alto de *homogeneidad empírica (consistencia interna), calculada (correlaciones)* no es garantía de *homogeneidad conceptual*. Cuando decimos que todos los ítems *miden lo mismo*, que *son homogéneos*, porque la fiabilidad es alta, lo que realmente queremos decir es que *las respuestas están de hecho relacionadas* pero no que los ítems (las preguntas) estén bien redactadas en torno a un mismo constructo o rasgo claramente definido. Hace falta también una *evaluación cualitativa, conceptual* de los ítems para poder afirmar que todos los ítems *miden lo mismo, expresan el mismo rasgo* tal como lo hemos concebido. Además varios subconjuntos de ítems muy relacionados entre sí pero marginalmente relacionados con otros subconjuntos de ítems pueden dar un coeficiente de fiabilidad alto en todo el instrumento y sin embargo un análisis conceptual de estos subconjuntos (más otros análisis estadísticos, como el *análisis factorial*) nos pueden llevar a la conclusión de que los subconjuntos miden rasgos suficientemente distintos como para que sea cuestionable sumarlos en un total único. *Consistencia interna* (tal como la cuantifican estos coeficientes) y *unidimensionalidad* son conceptos distintos, por eso decimos que un coeficiente alto de fiabilidad es un *apoyo pero no una prueba* de que el conjunto de ítems que componen el instrumento mide un único rasgo bien conceptualizado.

9.1.2. Fiabilidad y número de ítems

El coeficiente de fiabilidad aumenta al aumentar el número de ítems; ¿Quiere esto decir que los tests más largos son *más homogéneos*, que sus ítems miden con más claridad el mismo rasgo? Obviamente no; los ítems no están más relacionados entre sí por el mero hecho de ser más en número; el mismo Cronbach lo expresaba así: *un galón de leche no es más homogéneo que un vaso de leche*; un test no es *más homogéneo* por el mero hecho de ser más largo. El que al aumentar el número de ítems aumente la fiabilidad se debe, al menos en parte, a un mero mecanismo estadístico: cuando

[5] Sobre los *usos y abusos* del coeficiente α puede verse Schmitt (1996).

aumenta el número de ítems (con tal de que estén mínimamente relacionados entre sí) la suma de las covarianzas entre los ítems (numerador de la fórmula 4) aumenta proporcionalmente más que la varianza de los totales (denominador de la fórmula 4). Una fiabilidad alta se puede obtener con muchos ítems con relaciones bajas entre sí, e incluso con algunas negativas; y puede suceder también que (como ya hemos indicado) dos (o más) *bloques de ítems* con claras correlaciones entre los ítems dentro de cada bloque, pero con poca o nula relación con los ítems del *otro bloque* den para todo el test un coeficiente alto de fiabilidad. En este caso la *homogeneidad* del conjunto, y la interpretación de las puntuaciones como si expresaran *un único rasgo bien definido* puede ser cuestionable.

Por lo tanto:

a) No se debe buscar una fiabilidad alta aumentando sin más el número de ítems, sin pensar bien si son realmente *válidos* para expresar sin confusión el rasgo que deseamos medir. Una fiabilidad alta no es un indicador *cuasi automático* de la calidad de un test, sobre todo si es muy largo; hace falta siempre una *evaluación conceptual* de los ítems (además de verificar empíricamente su correlación con el total del instrumento).

b) Con frecuencia con un conjunto menor de ítems se puede conseguir una fiabilidad semejante o no mucho más baja que si utilizamos todos los ítems seleccionados en primer lugar, y varios subconjuntos de ítems pueden tener coeficientes de fiabilidad muy parecidos.

c) La fiabilidad también sube al aumentar el número de respuestas de los ítems (esto es más claro si pasamos de dos a tres o más respuestas); con un número menor de ítems pero con más respuestas se puede conseguir una fiabilidad semejante a la que conseguiríamos con más ítems y menos respuestas.

No hay que olvidar nunca que la *validez* es más importante que la fiabilidad; lo que más importa en primer lugar es que los ítems reflejen bien el rasgo que se desea medir.

9.1.3. Fiabilidad y simplicidad o complejidad del rasgo medido

Un coeficiente alto puede estar indicando que los ítems tienen homogeneidad conceptual, pero porque son *excesivamente repetitivos*, porque estamos midiendo un constructo o rasgo definido de manera muy limitada. Con pocos ítems muy repetitivos obtenemos con facilidad una fiabilidad alta.

Una *definición muy simple* de un rasgo no es necesariamente una mala característica cuando se trata hacer un instrumento de medición (puede ser incluso preferible) pero hay que tener en cuenta esta simplicidad de la concepción del rasgo en la interpretación, y más teniendo en cuenta que los *nombres* con que designamos a instrumentos y rasgos suelen ser muy genéricos *(autoestima, motivación, asertividad)* y la interpretación no debe hacerse en función del *nombre* del instrumento sino del contenido de los ítems que lo componen. Los nombres breves son cómodos, pero con frecuencia requieren alguna explicación adicional.

Un ejemplo claro y frecuente de un rasgo que a veces se mide de manera muy simple y otras de manera más compleja es la *autoestima*. Se puede preparar un instrumento de *autoestima general*, que incluirá múltiples aspectos (académico, social, familiar, etc.), o se puede construir un instrumento para medir la autoestima en un sentido muy restringido, como sería la *autoestima académica*.

También se pueden construir instrumentos *pluridimensionales:* se mide un rasgo complejo con todos los ítems del instrumento, y con una definición más bien genérica pero que *tiene sentido* (autoestima, asertividad, etc.) pero que a su vez se puede descomponer en *subescalas más específicas;* la fiabilidad puede calcularse tanto en todo el instrumento como en las subescalas que miden aspectos más simples.

9.2. El error típico de la medida

Una *utilidad importante* de los coeficientes de fiabilidad puede estar no en la magnitud misma de estos coeficientes, sino en los cálculos posteriores que podemos hacer a partir de los mismos. Uno

de estos cálculos es el del *error típico de la medida.* Es de especial utilidad cuando se van a hacer *interpretaciones individuales,* sobre todo si se derivan consecuencias importantes para los sujetos (aprobar, ser seleccionado para un puesto de trabajo, etc.), y con más razón si se juzga que la fiabilidad del instrumento dista de ser óptima. Ya hemos indicado en otro lugar que una fiabilidad alta es importante cuando los resultados (de un test) van a influir en la toma de decisiones sobre los sujetos (y el *aprobar* o *suspender* a un sujeto es una decisión importante).

9.2.1. Concepto y fórmula del error típico

El *error típico de la medida* viene a ser la *desviación típica de las puntuaciones individuales*, e indica el margen de error o *variación probable* de las puntuaciones individuales. En términos informales podemos decir que el error típico nos indica el *margen de oscilación probable* de las puntuaciones de una ocasión a otra o entre pruebas hipotéticamente iguales o semejantes. Nos puede servir para *relativizar* los resultados individuales.

Vamos a pensar en un ejemplo sencillo, un *examen tipo test.* Cada alumno tiene un resultado, su número de respuestas correctas.

Si cada alumno hubiera respondido a un número indefinido de exámenes, no hubiera obtenido en todos exactamente el mismo resultado; sus posibles resultados se hubieran distribuido según la *distribución normal* (figura 1).

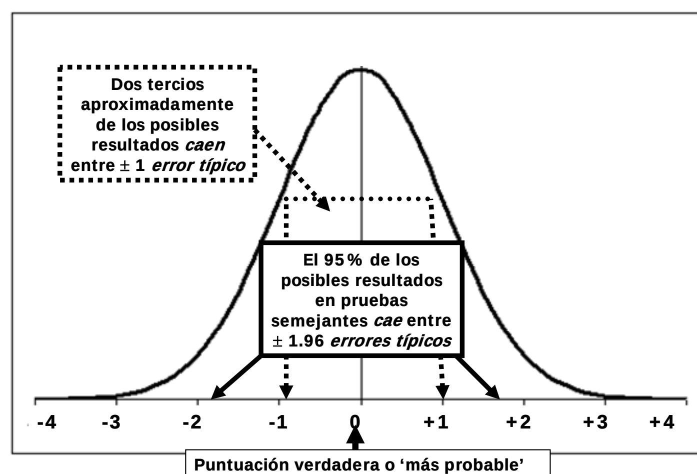


Figura 1

Esta distribución hubiera tenido su media y su desviación típica o *error típico de la medición*. Podemos suponer que la puntuación de hecho obtenida es la media de la distribución (aunque esto no es así exactamente, como veremos después al tratar de las *puntuaciones verdaderas*).

El *error típico de la medición* se calcula a partir del coeficiente de fiabilidad, y en muchos casos el *mejor uso* del coeficiente de fiabilidad es utilizarlo para calcular el *error típico,* (por ejemplo en *exámenes* o en cualquier test) cuando interese situar a cada uno en su *banda de posibles probables resultados*. Esta *banda de posibles resultados* será más *estrech*a (con un error típico menor) cuando la fiabilidad sea alta, y será más amplia cuando baje la fiabilidad. Una baja fiabilidad de un instrumento puede quedar neutralizada si utilizamos el error típico en la interpretación de las puntuaciones individuales.

La fórmula del error típico podemos derivarla con facilidad de las fórmulas [2] y [3].

De la fórmula [2] podemos despejar la varianza verdadera: $\sigma_v^2 = \sigma_t^2 - \sigma_e^2$

y substituyendo esta expresión de σ_v^2 en [3]: $$r_{11} = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2} = 1 - \frac{\sigma_e^2}{\sigma_t^2}$$

de donde $\frac{\sigma_e^2}{\sigma_t^2} = 1 - r_{11}$ y despejando σ_e tenemos que

error típico $\sigma_e = \sigma_t \sqrt{1 - r_{11}}$ [10]

Esta es la fórmula de la *desviación típica de los errores de medición*, denominada *error típico de la medida* o *de las puntuaciones individuales*. Se calcula a partir de la desviación típica (de los totales del test) y del coeficiente de fiabilidad calculados en la muestra. Si un sujeto hubiera respondido a una serie de *tests paralelos* semejantes, el error típico sería la desviación típica obtenida en esa serie de tests. Se interpreta como cualquier desviación típica e indica la variabilidad probable de las puntuaciones obtenidas, observadas.

El error típico es directamente proporcional al número de ítems y en el caso de los tests con respuestas 1 ó 0 (como en las pruebas objetivas) un cálculo rápido (y aproximado) es el dado en la fórmula [11][6]:

error típico $\sigma_e = .43\sqrt{k}$ [11]

Aquí hay que hacer una observación importante. Este error típico se aplica en principio *a todos los sujetos por igual;* hay *un* error típico que indica la *oscilación probable* de cada puntuación. Esto no es así exactamente. Pensemos en un examen: el alumno que *sabe todo,* en exámenes semejantes seguiría sabiendo todo, y el alumno que *no sabe nada,* en exámenes semejantes seguiría sin saber nada: la oscilación probable en los extremos es menor que en el centro de la distribución. Ésta es una limitación de esta medida del *error* probable individual. Aun así es la medida más utilizada aunque hay otras. Si la distribución es normal (o *aproximadamente* normal) y las puntuaciones máximas y mínimas obtenidas no son las máximas o mínimas posibles (la amplitud real no es igual a la amplitud máxima posible), éste *error típico de la medida* es más o menos uniforme a lo largo de toda la escala de puntuaciones.

Aquí nos limitamos a exponer el error típico habitual, el que se utiliza normalmente y que tiene aplicaciones muy específicas, pero en situaciones aplicadas (como en exámenes) sí conviene caer en la cuenta de que la posible variabilidad individual tiende a ser menor en los extremos de la distribución.

9.2.2. Las puntuaciones *verdaderas*

Un punto importante para el cálculo e interpretación del error típico es que el *centro de la distribución* de los posibles resultados no es para cada sujeto la puntuación que ha obtenido. Si un sujeto obtiene una puntuación de 120 y el error típico es de $\sigma_e = 4.47$, no podemos concluir que hay un 68% de probabilidades (aproximadamente, es la proporción de casos que suelen darse entre $\pm 1\sigma$) de que su verdadera puntuación está entre 120 ± 4.47. El *centro* de la distribución no es en este caso la puntuación obtenida, sino la denominada *puntuación verdadera* (X_V) que se puede estimar mediante esta fórmula:

Estimación de la puntuación verdadera: $X_v = [(X - \overline{X})(r_{11})] + \overline{X}$ [12]

En el caso anterior si $\overline{X} = 100$ y $r_{11} = .80$, la estimación de la puntuación verdadera de un sujeto que tuviera una puntuación de X = 120, sería [(120 - 100)(.80)] + 100 = 116. Si la fiabilidad es igual a 1, la puntuación obtenida es también la que aquí denominamos verdadera.

---

6 Puede verse explicado en Gardner (1970) y en Burton (2004). Hay varias fórmulas que permiten cálculos aproximados del error típico, del coeficiente de fiabilidad y de otros estadísticos que pueden ser útiles en un momento dado (por ejemplo, y entre otros, Saupe, 1961; McMorris, 1972).

Siguiendo con el mismo ejemplo, de un sujeto con X =120 y una puntuación verdadera de 116, podemos decir que sus posibles resultados en ese test (con un 5% de probabilidades de equivocarnos) están entre 116 ± (1.96 *errores típicos);* en este caso entre 116 ± (1.96)(4.47) o entre 107 y 125.

Estas *puntuaciones verdaderas* tienden a ser menores que las obtenidas cuando estas son superiores a la media, y mayores cuando son inferiores a la media. No debemos entender esta puntuación *verdadera* (aunque éste sea el término utilizado) como expresión de una *verdad absoluta,* que nos dice exactamente lo que *vale* o *sabe* una persona en aquello en la que la hemos medido. Hay que entender más bien esta *puntuación verdadera* como la *puntuación más probable* que un sujeto hubiera obtenido si le hubiéramos medido repetidas veces en el mismo rasgo y con el mismo instrumento.

Las puntuaciones *verdaderas* y las puntuaciones *observadas* tienen una correlación perfecta (el orden de los sujetos es el mismo con las dos puntuaciones) por lo que el cálculo de estas *puntuaciones verdaderas* no tiene siempre una especial utilidad práctica; sí puede tenerla cuando se desea precisamente utilizar el error típico para precisar *con mayor rigor* y exactitud *entre qué límites* o *banda de resultados probables* se encuentra la verdadera puntuación, como tratamos en el apartado siguiente.

9.2.3. Los *intervalos de confianza de las puntuaciones individuales*

Como el error típico se interpreta como una *desviación típica,* si el error típico es de 4.47, hay un 68% de probabilidades de que la verdadera puntuación estaría entre 116 ± 4.47 (la *puntuación verdadera más-menos un error típico;* es la proporción de casos que caen en la distribución normal entre la media más una desviación típica y la media menos una desviación típica, como se representa en la figura 1).

Podemos establecer *intervalos de confianza* con mayor seguridad, y así podríamos decir, con un 95% de probabilidades de acertar (z = 1.96) que la puntuación verdadera se encuentra entre 116 ± $1.96\sigma_e$ y en nuestro ejemplo entre 116 ± (1.96)(4.47) o entre 116 ± 8.76 (es decir, entre 107 y 125).

El error típico nos sirve para *relativizar* las puntuaciones obtenidas, y más que pensar en una puntuación concreta, la obtenida por cada sujeto, podemos pensar en una *banda de posibles puntuaciones*.

La *puntuación verdadera* exacta de cada sujeto (la que hubiera obtenido respondiendo a todos los ítems del *universo* e ítems o a muchas pruebas paralelas) no la sabemos, pero sí podemos estimar *entre qué límites se encuentra*, y esto puede ser de utilidad práctica en muchas ocasiones (como cuando en un examen hay una *puntuación mínima* para el apto: sumando a los que están en el *límite* un error típico, o *margen de oscilación probable,* algunos quizás superen holgadamente ese límite; al menos hay un criterio razonablemente objetivo, justificable y común para todos).

El error típico lo aplicamos a todos los sujetos por igual, y es ésta una limitación de esta medida de *error*. Frecuentemente el error probable es menor en los extremos de la distribución (simplificando la explicación: el que *sabe todo* y el que *no sabe nada* obtendrían los mismos resultados en pruebas sucesivas). Aun así si la distribución se aproxima a la distribución normal y las puntuaciones extremas obtenidas no son las máximas posibles (la amplitud real no es igual a la amplitud máxima), el error típico de la medida es uniforme, más o menos, a lo largo de toda la escala de puntuaciones.

9.3. Coeficientes de correlación corregidos por atenuación

En buena medida la *utilidad de los coeficientes de fiabilidad* está en los cálculos adicionales que podemos hacer. Posiblemente el más importante, y de utilidad práctica, es el del *error típico* de la medida que ya hemos visto.

Otra utilidad de estos coeficientes es que nos permiten calcular el valor de un coeficiente de correlación entre dos variables *corregido por atenuación*.

La correlación calculada entre dos variables queda siempre disminuida, *atenuada*, por culpa de los *errores de medición*, es decir, por su no perfecta fiabilidad. La *verdadera relación* es la que tendríamos si nuestros instrumentos midieran sin error. Esta correlación *corregida por atenuación* es la que hubiéramos obtenido si hubiésemos podido suprimir los errores de medición en las dos variables (o al menos en una de las dos; no siempre conocemos la fiabilidad de las dos variables).

Conociendo la fiabilidad de las dos variables podemos *estimar* la verdadera relación mediante esta fórmula:

$$r_{xy}\ \textit{(corregida)} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}} \qquad [13]$$

En esta fórmula r_{xy} es el coeficiente de correlación obtenido entre dos variables, X e Y, y r_{xx} y r_{yy} son los coeficientes de fiabilidad de cada variable; si conocemos solamente la fiabilidad de una de las dos variables, en el denominador tendremos solamente la raíz cuadrada de la fiabilidad conocida.

Por ejemplo si entre dos tests o escalas tenemos una correlación de .30 y los coeficientes de fiabilidad de los dos tests son .50 y .70, la correlación estimada corregida por atenuación sería:

$$r_{xy}\ \textit{(corregida)} = \frac{.30}{\sqrt{.50}\sqrt{.70}} = .507$$

Vemos que la correlación sube apreciablemente; y expresa la relación entre las dos variables independientemente de los errores de medición de los instrumentos utilizados.

Sobre estas *estimaciones* de la correlación entre dos variables (entre las *verdaderas* puntuaciones de X e Y, sin errores de medición) ya se han hecho una serie de observaciones al tratar sobre los *coeficientes de correlación* (ése es el contexto apropiado); conviene tener en cuenta esas observaciones (que no repetimos aquí) sobre 1º en qué condiciones se debe utilizar esta fórmula de *corrección por atenuación,* 2º en qué situaciones es más útil y 3º otras fórmulas distintas de corrección por atenuación. Conviene repasar estas observaciones antes de aplicar estas fórmulas[7].

10. Cuando tenemos un coeficiente de fiabilidad bajo

Un coeficiente de fiabilidad *bajo* no indica necesariamente que el instrumento es malo y que no es posible utilizarlo. También puede suceder que haya una razonable *homogeneidad conceptual* en la formulación de los ítems, y esto se procura siempre, y que esta homogeneidad no se refleje en un coeficiente alto de fiabilidad. En cualquier caso con un coeficiente de fiabilidad bajo y si se van a tomar decisiones sobre los sujetos (una decisión puede ser dar un informe) sí conviene incorporar el *error típico* a la interpretación.

Ahora nos interesa examinar *de dónde* puede venir un bajo coeficiente de fiabilidad.

10.1. Inadecuada formulación de los ítems

Puede ser que los sujetos *entiendan los ítems de una manera distinta a como lo pretende el autor del instrumento*. Un *a veces me gustaría marcharme de casa* podría significar para algunos *me gusta viajar, etc.* y en este caso las respuestas no *serían coherentes* con *el significado pretendido* por el constructor del instrumento (*me siento mal en casa)*. La *coherencia conceptual prevista* la comprobamos con la *coherencia que de hecho encontramos en las respuestas*. En el análisis de ítems, al construir un instrumento, podemos comprobar si los sujetos que responden, parecen entender la formulación con el significado previsto; en caso contrario tendremos que eliminarlos o reformularlos.

---

7 Una buena exposición de los efectos de la baja fiabilidad en los coeficientes de correlación y de la *corrección por atenuación* puede verse en Osborne (2003).

10.2. Homogeneidad de la muestra

Podemos encontrarnos con una *homogeneidad conceptual clara* y una fiabilidad muy baja, y una causa importante puede estar en que *apenas hay diferencias entre los sujetos* (todos o casi todos se encuentran bien con sus padres, no les gustaría marcharse de casa, etc.). Si no hay diferencias tampoco habrá *relación clara y verificada* entre las respuestas (sin diferencias entre los sujetos los *coeficientes de correlación* entre los ítems son muy bajos). Por eso *la fiabilidad es mayor con muestras heterogéneas,* en las que hay mayores diferencias en las respuestas. Con una muestra más variada (o simplemente mayor, donde es más probable que haya sujetos muy diferentes) podemos encontrar una fiabilidad alta. De todas maneras con una fiabilidad baja que no se deba a la mala calidad del instrumento sino a la homogeneidad de la muestra, seguiremos *clasificando mal (diferenciando, midiendo* mal) a los sujetos de esa muestra.

Además de la *homogeneidad conceptual* propia de los ítems que hipotéticamente definen un rasgo, es necesario comprobar la relación *de hecho* entre las respuestas (y esto indica la fiabilidad) porque esta relación supone que podemos *diferenciar a los sujetos* y ésta es la única manera de *medir, clasificar, comparar,* etc. Cuando construimos una escala o test, estamos construyendo un *instrumento de medición* con el que podamos diferenciar adecuadamente a los sujetos según *tengan más o menos* del rasgo que pretendemos medir.

10.3. Definición compleja del rasgo medido

Por supuesto una fiabilidad baja, sobre todo si la obtenemos con una muestra razonablemente heterogénea, puede significar una concepción del rasgo *excesivamente compleja* o una construcción deficiente del instrumento. Aun así podemos encontrar coeficientes bajos en tests reconocidos como buenos porque miden rasgos definidos con un grado grande de complejidad.

Rasgos definidos de manera *compleja* o *muy genérica* pueden tener ítems poco relacionados entre sí y consecuentemente tendremos una fiabilidad baja aunque esté presente la *unidad conceptual* pretendida por el autor. La consecuencia es que en estos casos se puede llegar a una misma puntuación total por caminos distintos, y esto hay que asumirlo en la interpretación. En cualquier caso la fiabilidad debería estar dentro de unos mínimos aceptables para poder afirmar que estamos *midiendo, diferenciando* a los sujetos según posean más o menos del *rasgo* que supuestamente medimos[8].

Cuando la fiabilidad es baja, observando la redacción de los ítems y cómo se relacionan entre sí, podemos llegar a la conclusión que es preferible una concepción más simple del rasgo, sin mezclar ideas relacionadas pero no lo suficiente, o dividir el instrumento en dos (o más) instrumentos y medir aspectos distintos por separado con instrumentos distintos.

10.4. Utilidad del error típico cuando la fiabilidad es baja

Una *valoración racional* del coeficiente de fiabilidad tendrá en cuenta tanto la homogeneidad de la muestra como la complejidad del instrumento, y en cualquier caso con coeficientes bajos siempre es conveniente utilizar el *error típico* en la interpretación de los resultados individuales. Cuando se trata de tomar decisiones sobre sujetos, o de dar un informe de cierta importancia (por ejemplo en un psicodiagnóstico) y la fiabilidad del instrumento es baja, es cuando puede ser de especial utilidad (e incluso de *responsabilidad ética)* no limitarse a informar con una puntuación o resultado muy preciso, sino con *una banda de puntuaciones probables;* esta banda o límites probables de la puntuación será mayor cuando el error típico sea mayor (y la fiabilidad más baja).

---

[8] Un tratamiento más extenso de la *fiabilidad* y de la *unidimensionalidad* de los tests puede verse en Morales (2006, cap. 9 y 10).

11. La fiabilidad en exámenes y pruebas escolares

En primer lugar recordemos que es relativamente frecuente calcular la fiabilidad de las pruebas *tipo test* (estos cálculos, y otros, suelen estar programados), pero también se puede calcular la fiabilidad de un examen compuesto por unas pocas preguntas de respuesta abierta, con tal de que en todas las preguntas se utilice la misma clave de corrección. Las fórmulas adecuadas las veremos después; en las pruebas cuyos ítems puntúan 1 ó 0 (lo habitual con pruebas objetivas) se utiliza alguna de las fórmulas de Kuder-Richardson, y cuando las puntuaciones son continuas (por ejemplo de 0 a 4 o algo similar) se utiliza el coeficiente α de Cronbach.

Cuando se trata de *exámenes escolares* el coeficiente de fiabilidad puede presentar problemas específicos de interpretación. No hay que olvidar que la psicometría clásica trata de las diferencias individuales en medidas psicológicas que parten al menos de dos supuestos:

a) Todos los componentes (ítems) del test miden el mismo rasgo.
b) Los sujetos son distintos en el rasgo que queremos medir.

Estos dos supuestos no son aplicables siempre y automáticamente a los diversos tipos de exámenes y pruebas escolares. En estas pruebas los coeficientes de fiabilidad pueden dar información útil, pero hay que tener cuidado en la interpretación.

Es importante pensar en la fiabilidad de los exámenes porque se interpreta y utiliza habitualmente como un *control de calidad,* y se estima que *siempre es bueno* que un test de conocimientos (como un examen *tipo test*) tenga una fiabilidad alta. En el caso de los exámenes esto puede ser discutible (aunque no en todas las situaciones) y conviene hacer algunas matizaciones.

11.1. Fiabilidad y validez

En primer lugar la característica más importante de una prueba escolar (como de cualquier instrumento de medición) no es la *fiabilidad psicométrica,* sino la *validez:* una prueba de evaluación o cualquier examen es bueno si comprueba los objetivos deseados (y comunicados previamente), si condiciona en el alumno un estudio inteligente. Con una prueba objetiva se puede conseguir fácilmente una fiabilidad muy alta, pero se pueden estar comprobando meros conocimientos de memoria cuando quizás el objetivo pretendido era (o debería ser) de comprensión. La *validez* es por lo tanto la primera consideración para *evaluar la evaluación*: en principio un instrumento es válido si mide lo que decimos que mide.

11.2. Fiabilidad y diferencias entre los sujetos

Por lo que respecta a la fiabilidad, hay que tener en cuenta que en última instancia la fiabilidad expresa la *capacidad diferenciadora* de un test, y esto es en principio deseable cuando se trata precisamente de diferenciar. Si un *test de inteligencia* no diferencia adecuadamente a los más y a los menos inteligentes (y lo mismo diríamos de cualquier otra capacidad o rasgo psicológico) sencillamente no nos sirve. En definitiva en estos casos *medir* es *diferenciar.* Por eso en todo tipo de *tests psicológicos,* escalas de actitudes, etc., una fiabilidad alta es una característica deseable. Entendiendo bien que la fiabilidad no es una característica de un test (aunque ésta sea la expresión habitual) sino de un conjunto de puntuaciones que quedan mejor o peor diferenciadas.

Si pensamos en los tests escolares de conocimientos, podemos preguntarnos si las diferencias son deseables, si es verdad que un test que distingue, matiza y establece diferencias nítidas entre los alumnos implica que tenemos un buen test y, sobre todo, unos buenos resultados. Obviamente esto puede ser discutible.

Una fiabilidad baja en un *examen* puede provenir de cualquiera de estas dos circunstancias: sujetos muy igualados o preguntas muy distintas (el saber unas no implica saber otras).

a) *La clase está muy igualada,* apenas hay diferencias pronunciadas o *sistemáticas* entre los alumnos. No se puede clasificar bien a los *inclasificables*. Que esto sea bueno o malo deberá juzgarlo el profesor. En un test sencillo de *objetivos mínimos* un buen resultado es que *todos sepan todo,* y en este caso la fiabilidad sería igual a *cero*. Lo mismo puede suceder con un test más difícil, sobre todo en grupos pequeños, en los que todos los alumnos tienen un rendimiento alto.

b) *Las preguntas son muy distintas* y el saber unas cosas no implica saber otras, *no hay homogeneidad en los ítems* ni se pretende. Esta situación no suele ser la más frecuente en los tests escolares más convencionales, pero si no hay homogeneidad en las preguntas de un test (porque se preguntan cosas muy distintas o de manera muy distinta) y el saber unas cosas no implica saber otras, entonces lógicamente bajará la fiabilidad de todo el test (debido a la poca relación entre unas y otras preguntas o ejercicios).

En un *examen final* más o menos *largo,* donde *hay de todo,* fácil y difícil, en una clase relativamente numerosa, en la que hay alumnos más y menos aventajados, una fiabilidad alta en una *prueba objetiva* nos indicará que detectamos bien diferencias que de hecho existen y que además son *legítimas* o al menos esperables. Cuando *todos saben todo* en un examen de esas características, esto puede significar que estamos igualando a la clase por su nivel más bajo y que el profesor *no da juego* a los más capaces.

11.3. Fiabilidad y calificación

También hay que pensar que una fiabilidad alta indica en principio *diferencias consistentes* entre los alumnos, pero no indica necesariamente que los de puntuación más baja no lleguen al nivel del apto. Si todos los alumnos están en la parte alta de la distribución pero bien diferenciados, la fiabilidad será alta; en este caso los que saben menos pueden saber lo suficiente; y también puede suceder lo contrario, que los que saben más que los demás no sepan lo suficiente.

Lo que sí parece claro es que una fiabilidad alta es deseable en todo instrumento de medida cuya función y utilidad está precisamente en que nos permite conocer si un sujeto *tiene mucho o poco* del rasgo que estamos midiendo y *además* nos interesa *diferenciar* a unos sujetos de otros, o al menos es razonable esperar diferencias claras entre los sujetos (como ya se ha indicado en 11.2).

Lo que sí puede ser siempre de utilidad en cualquier tipo de examen es calcular y utilizar el *error típico de la medida* o de las puntuaciones obtenidas (para lo cual necesitamos el coeficiente de fiabilidad)[9], porque nos indica la *banda probable de resultados* en la que se encuentra cada alumno, y esta *banda,* aunque sea más imprecisa, refleja mejor que *un número exacto* de respuestas correctas *por dónde se encuentra* cada uno. En lenguaje coloquial podríamos decir que el error típico expresa el *margen de mala o buena suerte* del alumno ante unas preguntas concretas, y puede ayudar a relativizar una *mera suma* de respuestas correctas. Si establecemos previamente una puntuación *de corte* para situar el aprobado, el sumar, por ejemplo, un error típico a los alumnos que están en el *límite del apto* puede ser una buena práctica (como ya se ha indicado en el apartado 9.2.3)[10].

En los tests o pruebas objetivas de *criterio,* en los que hay una *puntuación de corte* para distinguir al *apto* del *no apto,* la distribución deja de ser normal (una condición de los modelos

---

[9] *I am convinced that the standard error of measurement… is the most important single piece of information to report regarding an instrument, and not a coefficient* (Cronbach y Shavelson, 2004).Ya hemos indicado que un cálculo aproximado y rápido del *error típico de la media* es $.43\sqrt{k}$ donde k es el número de ítems (Burton, 2004).

[10] Si en un examen tipo test sumamos a los que están justo debajo del límite propuesto para el aprobado *dos erres típicos* nos ponemos prácticamente en el límite máximo probable al que hubiera llegado ese alumno.

teóricos de las fórmulas) y disponemos de *estimaciones* de la fiabilidad más idóneas que las fórmulas usuales (fórmula 24).

12. Fórmulas de los coeficientes de *consistencia interna*

Las fórmulas del coeficiente de fiabilidad son muchas, aquí exponemos las más utilizadas. Podemos dividirlas en dos grupos:

1) Fórmulas que se basan en la partición del test en dos mitades

2) Fórmulas en las que se utiliza información de todos los ítems, como las de Kuder-Richardson y Cronbach.

En cada uno de los apartados se incluyen otras fórmulas relacionadas o derivadas; también exponemos otras formulas de interés, como las fórmulas que relacionan la fiabilidad con el número de ítems.

12.1. *Fórmulas basadas en la partición del test en dos mitades*

12.1.1. Cómo dividir un test en dos mitades

1. Como cualquier test puede dividirse en muchas dos mitades, puede haber muchos coeficientes distintos de fiabilidad. El resultado es sólo una *estimación* que puede infravalorar o supervalorar la fiabilidad. Es habitual la práctica de dividir el test en ítems *pares* e *impares,* pero puede dividirse en dos mitades cualesquiera. Cada mitad debe tener el mismo número de ítems o muy parecido.

2. Si al dividir el test en dos mitades *emparejemos* los ítems según sus contenidos *(matching),* de manera que cada mitad del test conste de ítems muy parecidos, obtendremos una *estimación más alta* y preferible de la fiabilidad.

3. Cuando la mitad (o *casi* la mitad) de los ítems son *positivos* y la otra mitad son *negativos* (*favorables* o *desfavorables* al rasgo medido, con distinta clave de corrección), es útil que las dos mitades estén compuestas una por los ítems positivos y otra por los negativos. En este caso la correlación entre los dos tipos de ítems es muy informativa en sí misma, aunque no se calcule después la fiabilidad por este procedimiento. Una correlación entre los dos subtests en torno a .50 o mayor indica *suficiente coherencia* entre los dos tipos de ítems, y que no se manifiesta de modo apreciable la *aquiescencia* o tendencia a mostrar acuerdo (o responder *sí)* a ítems que expresan ideas contradictorias.

12.1.2. Fórmulas

De estas fórmulas la primera y más clásica es la de Spearman-Brown; ambos autores derivaron las mismas fórmulas de manera independiente en 1910 (la fórmula básica de estos autores es la 19, de la que se derivan la 13, la 21 y la 22). Esta es la fórmula que se conoce habitualmente como procedimiento de las *dos mitades* (vamos a ver que además hay otras fórmulas) y no suele faltar en ningún texto cuando se trata de la fiabilidad; es la fórmula [14].

$$r_{11} = \frac{2r_{12}}{1+r_{12}} \qquad [14]$$

Fórmula de Spearman-Brown

r_{12} = correlación entre las dos mitades del test. El test se divide en dos mitades y se calcula la correlación entre ambas como si se tratara de dos tests.

1. La correlación entre las dos mitades es la fiabilidad de una de las dos (pruebas paralelas); con esta fórmula se calcula la fiabilidad de todo el test. Observando la fórmula puede verse que si r_{12} = 1, también tendremos que r_{11} = 1.

2. La fórmula [14] supone que las dos mitades tienen medias y varianzas idénticas; estos presupuestos no suelen cumplirse nunca, y de hecho con esta fórmula se sobrestima la fiabilidad; por lo que esta fórmula está desaconsejada (a pesar de su uso habitual); la fórmula de las dos mitades preferible es la conocida como *dos mitades alpha* ($r_{2\alpha}$) [15][11]:

$$r_{2\alpha} = \frac{4\,r_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + 2\,r_{12}\sigma_1\sigma_2} \qquad [15]$$

En esta fórmula *entran* también, además de la correlación de las dos mitades, las desviaciones típicas de cada mitad.

3. Otras fórmulas basadas en la partición de un test en dos mitades, y que suelen encontrarse en algunos textos, son la [16] y la [17], que no requieren el cálculo de la correlación entre las dos mitades; de todas maneras en estos casos (partición del test en dos mitades) es siempre preferible la fórmula [15].

$$r_{11} = 2\left(1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_t^2}\right) \qquad [16]$$

Fórmula de Flanagan[12]

σ_1^2 y σ_2^2 son las varianzas de las dos mitades, σ_t^2 es la varianza de todo el test

$$r_{11} = 1 - \frac{\sigma_d^2}{\sigma_t^2} \qquad [17]$$

σ_d^2 = Es la varianza de la diferencia entre las dos mitades. Cada sujeto tiene dos puntuaciones, una en cada mitad: a cada sujeto se le restan ambas puntuaciones y se calcula la varianza de estas diferencias.

Como cálculo aproximado y rápido de la fiabilidad la fórmula más cómoda es la formula [19] que veremos después, pero sólo si los ítems son dicotómicos (puntúan 1 ó 0). Las fórmulas preferibles, y que deben utilizarse habitualmente, son las de Kuder-Richardson [17] (para ítems dicotómicos) y el α de Cronbach [20], y son las que suelen venir programadas en los programas de ordenador.

12.2. Fórmulas de Kuder-Richardson y α de Cronbach

Se trata de las fórmulas de *consistencia interna* que hemos justificado anteriormente con mayor amplitud; son las más utilizadas.

a) Son métodos en principio preferibles porque con los métodos de las *dos mitades* cabe dividir un test en muchas dos mitades con que las que podemos obtener distintos valores del coeficiente de fiabilidad. El resultado que nos dan las fórmulas de Kuder-Richardson y Cronbach equivale a la fiabilidad media que obtendríamos dividiendo un test en todas sus posibles dos mitades; obtenemos un único coeficiente que es una estimación *más segura.*

---

[11] Esta fórmula la aconsejan buenos autores (incluido el mismo Cronbach). La importancia del cálculo de la fiabilidad por el procedimiento de las *dos mitades* es sobre todo histórica; el método de las *pruebas paralelas* (dos pruebas en vez de dos mitades) y el de *consistencia interna* (en el que cada ítem *funciona* como una prueba paralela) parten de la intuición original de las *dos mitades* de Spearman y Brown. Una crítica y valoración de estas fórmulas puede verse en Charter (2001).

[12] Esta fórmula también se conoce como *fórmula de Rulon* que es el primero que la expuso (en 1939) aunque Rulon se la atribuye a Flanagan (Traub, 1994).

b) En los modelos teóricos de donde parten estas fórmulas se supone que tanto las varianzas como las intercorrelaciones de los ítems son iguales; esto no suele suceder por lo que estas fórmulas tienden a dar una estimación de la fiabilidad algo baja.

c) Las fórmulas de Kuder-Richardson son válidas para ítems dicotómicos (0 ó 1), y el coeficiente α de Cronbach para ítems con repuestas continuas (más de dos repuestas).

$$r_{11} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma_t^2}\right) \qquad [18]$$

fórmula Kuder-Richardson 20
(para ítems dicotómicos)

k = número de ítems
Σpq = suma de las varianzas de los ítems
σ_t^2 = varianza de los totales

Como ya sabemos, p es la proporción de *unos* (aciertos, *síes*, la respuesta que se codifique con un 1) y q es la proporción de *ceros* (número de unos o de ceros dividido por el número de sujetos).

Con *ítems dicotómicos* ésta es la fórmula [18] que en principio debe utilizarse. Si se tienen calculadas las varianzas o desviaciones típicas de cada ítem, no es muy laboriosa.

Si el cálculo resulta laborioso y no se tiene ya programada la fórmula completa de la fiabilidad, hay otras alternativas más sencillas; la más utilizada es la fórmula Kuder-Richardson 21[13].

$$r_{11} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\overline{X}(k - \overline{X})}{k\sigma_t^2}\right) \qquad [19]$$

fórmula Kuder-Richardson 21

k es el número de ítems;
$\overline{X}$ y σ_t^2 son la media y varianza de los totales

1. Esta fórmula [19] se deriva de la anterior [18] si suponemos que todos los ítems tienen idéntica media. En este caso tendríamos que:

$$\Sigma pq = (k)\,(\overline{p})(\overline{q}) \qquad \overline{p} = \frac{\overline{X}}{N} \qquad y \qquad \overline{q} = 1 - \overline{p}$$

Haciendo las sustituciones oportunas en [18] llegamos a la fórmula [19].

2. Esta fórmula es sencilla y cómoda, porque solamente requiere el cálculo de la media y varianza de los totales. La suposición de que todos los ítems tienen idéntica media no suele cumplirse, por lo que esta fórmula da una estimación de la fiabilidad todavía más imprecisa. Es útil cuando no se necesita una estimación muy exacta; se utiliza frecuentemente para calcular la fiabilidad de las pruebas objetivas (exámenes, evaluaciones) hechas por el profesor y por lo menos indica *por dónde va* la fiabilidad; puede ser suficiente para calcular el *error típico* y *relativizar* los resultados individuales.

Existen otras aproximaciones de la fórmula Kuder-Richardson 20, pero es ésta la más utilizada.

Con *ítems continuos,* con más de una respuesta como los de las *escalas de actitudes,* la fórmula apropiada es la del coeficiente α de Cronbach que es una generalización de la Kuder-Richardson 20; es la fórmula [8] que ya vimos antes:

---

[13] El coeficiente de fiabilidad se calcula en el SPSS en la opción *analizar,* en *escalas.*

$$\alpha = \left(\frac{k}{k-1}\right)\left(1-\frac{\Sigma\sigma_i^2}{\sigma_t^2}\right) \quad [20]$$

k = número de ítems
$\Sigma\sigma_i^2$ es la suma de las varianzas de los ítems
σ_t^2 es la varianza de los totales.

α de Cronbach para ítems continuos

12.3. Fórmulas que ponen en relación la fiabilidad y el número de ítems

1. La fórmula [14] se deriva de esta otra, denominada *fórmula profética de Spearman-Brown:*

$$r_{kk} = \frac{k\bar{r}_{ij}}{1+(k-1)\bar{r}_{ij}} \quad [21]$$

r_{kk} = fiabilidad de un test compuesto por k ítems
$\bar{r}_{ij}$ = correlación media entre los ítems

En la fórmula [14] hemos supuesto que k =2 y $\bar{r}_{ij} = r_{12}$ De la fórmula anterior [21] se derivan otras dos especialmente útiles, y que se pueden utilizar aunque la fiabilidad no se calcule por el método de Spearman-Brown.

12.3.1. Cuánto aumenta la fiabilidad al aumentar el número de ítems

Disponemos de una fórmula que nos dice (siempre de manera aproximada) en cuánto aumentará la fiabilidad si aumentamos el número de ítems multiplicando el número de ítems inicial, que ya tenemos, por un factor *n*. Es en realidad una aplicación de la misma fórmula.

$$r_{nn} = \frac{nr_{11}}{1+(n-1)r_{11}} \quad [22]$$

r_{nn} = nuevo coeficiente de fiabilidad estimado si multiplicamos el número de ítems que tenemos por el factor n

r_{11} = coeficiente de fiabilidad conocido

n = factor por el que multiplicamos el número de ítems

Por ejemplo: tenemos una escala de actitudes de 10 ítems y una fiabilidad de .65. La fiabilidad nos parece baja y nos preguntamos cuál será el coeficiente de fiabilidad si multiplicamos el número de ítems (10) por 2 (n = 2) y llegamos así a 20 ítems (del *mismo estilo* que ya los que ya tenemos). Aplicando la fórmula anterior tendríamos:

$$r_{nn} = \frac{(2)(.65)}{1+(2-1)(.65)} = .788$$

multiplicando por 2 el número inicial de ítems llegaríamos a una fiabilidad en torno a .80

Si en la fórmula [22] hacemos n = 2, tendremos la fórmula [14]; r_{12} es la fiabilidad de una de las dos mitades, lo que nos dice la fórmula [14] es la fiabilidad del test entero (formado por las dos mitades)[14].

12.3.2. En cuánto debemos aumentar el número de ítems para alcanzar una determinada fiabilidad

Posiblemente es más útil la fórmula siguiente. Si tenemos una fiabilidad conocida (r_{11}) y queremos llegar a otra más alta (esperada, r_{nn}), ¿En cuántos ítems tendríamos que alargar el test? En este caso nos preguntamos por el valor de *n*, el factor por el que tenemos que multiplicar el número de ítems que ya tenemos.

[14] A partir de una fiabilidad obtenida con un número determinado de ítems puede verse en Morales, Urosa y Blanco (2003) una tabla con la fiabilidad que obtendríamos multiplicando el número inicial de ítems por un factor n.

$$n = \frac{r_{nn}(1- r_{11})}{r_{11}(1- r_{nn})} \qquad [23]$$

n = factor por el que debemos multiplicar el número de ítems para conseguir una determinada fiabilidad

r_{nn} = fiabilidad deseada

r_{11} = fiabilidad obtenida con el número original de ítems

Si, por ejemplo, con 8 ítems hemos conseguido una fiabilidad de .57 y deseamos llegar a una fiabilidad aproximada de r_{nn} = .75, ¿Por qué coeficiente *n* deberemos multiplicar nuestro número inicial de ítems?

$$n = \frac{.75(1-.57)}{.57(1-.75)} = 2.26;$$ necesitaremos (2.26)(8) = 18 ítems.

Naturalmente los nuevos ítems deben ser parecidos a los que ya tenemos. Si el número de ítems que necesitamos para alcanzar una fiabilidad aceptable es obviamente excesivo, posiblemente los contenidos del núcleo inicial de ítems no representan bien un rasgo definido con claridad (al menos para la población representada por *esa* muestra) y es preferible *intentar otra cosa.*

12.4. Estimación de la fiabilidad en una nueva muestra cuya varianza conocemos a partir de la varianza y fiabilidad calculadas en otra muestra

La fiabilidad hay que calcularla en cada muestra. Al obtener los datos con un test en una nueva muestra no se puede aducir la fiabilidad obtenida en otras muestras como prueba o garantía de que en la nueva muestra la fiabilidad será semejante[15]. En definitiva la fiabilidad indica en qué grado el test diferencia a unos sujetos de otros y esto depende de la heterogeneidad de la muestra; por lo tanto se puede ordenar bien a los sujetos de una muestra y no tan bien a los de otra muestra distinta en la que los sujetos estén más igualados. En nuevas muestras con una varianza menor, lo normal es que la fiabilidad baje.

Lo que sí se puede hacer es *estimar* la fiabilidad en una nueva muestra conociendo su desviación típica a partir de la fiabilidad obtenida en otra muestra de la que también conocemos la desviación típica (Guilford y Fruchter, 1973:420), bien entendido que se trata solamente de una *estimación*.

$$\mathbf{r_{nn}} = 1 - \frac{\sigma_{\mathbf{o}}^2(1-\mathbf{r_{oo}})}{\sigma_{\mathbf{n}}^2} \qquad [24]$$

r_{nn} = *fiabilidad estimada* en la nueva muestra

σ_o y r_{oo} = desviación típica y fiabilidad ya calculadas (observadas) en una muestra

σ_n = desviación típica en la nueva muestra (en la que deseamos *estimar* la fiabilidad)

Por ejemplo, si en una escala de actitudes hemos obtenido en una muestra una desviación típica de 6.86 y una fiabilidad de α = .78 ¿qué fiabilidad podemos esperar en otra muestra cuya desviación típica vemos que es 7.28?

Aplicando la fórmula anterior: *fiabilidad esperada* = $1 - \frac{6.68^2(1-.78)}{7.28^2}$ = .8147

De hecho la fiabilidad calculada en la nueva muestra (ejemplo real) es de 8.15, aunque no siempre obtenemos unas estimaciones tan ajustadas.

---

15 El obtener la fiabilidad en cada nueva muestra es una de las recomendaciones de la American Psychological Association (5ª edición, 2001).

12.5. La fiabilidad en los tests de criterio

Suelen denominarse *tests de criterio* los tests en los que se establece una *puntuación de corte* para separar al *apto* del *no apto*. En estos casos no se cumple la condición de normalidad, se reduce la dispersión de la muestra y la fiabilidad se calcula de otras maneras (pueden verse referencias en los comentarios bibliográficos). Un índice apropiado en estos casos es el de Livingston (Mehrens y Lehmann, 1984).

$$r_{cc} = \frac{r_{xx}\sigma^2 + (\overline{X} - C)^2}{\sigma^2 + (\overline{X} - C)^2} \qquad [25]$$

fórmula de Livingstone

$\overline{X}$ y σ^2 son la media y la varianza observadas;
C es la *puntuación de corte,*
r_{xx} es la fiabilidad usual (Kuder-Richardson o α de Cronbach).

Como puede verse en la fórmula, si la media es igual a la *puntuación de corte,* este índice equivale a los coeficientes normales de fiabilidad (en Black, 1999:292, puede verse también el *índice de discriminación* apropiado en los *tests de criterio).*

13. Resumen: concepto básico de la fiabilidad en cuanto *consistencia interna*

En el cuadro puesto a continuación tenemos un *resumen significativo* de lo que significa la fiabilidad en cuanto consistencia interna, cómo se interpreta y en qué condiciones tiende a ser mayor.

1. Cuando ponemos un test o una escala aun grupo de sujetos nos encontramos con diferencias inter-individuales. Estas diferencias o diversidad en sus puntuaciones totales las cuantificamos mediante la desviación típica (σ) o la varianza (σ^2).
2. Esta varianza (diferencias) se debe a las respuestas de los sujetos que pueden ser de dos tipos (fijándonos en los casos extremos; hay grados intermedios): coherentes (relacionadas) o incoherentes, por ejemplo:

	respuestas coherentes	*respuestas incoherentes*
En mi casa me siento mal	de acuerdo	en desacuerdo
A veces me gustaría marcharme de casa	de acuerdo	de acuerdo

3. La *incoherencia* aquí quiere decir que la respuesta no está en la dirección de las otras, tal como lo pretende el autor del instrumento (y esto por cualquier razón: pregunta ambigua, el que responde lo entiende de otra manera, etc.). Las respuestas coherentes son las respuestas relacionadas.

Diversidad (o *varianza)* total =

diversidad debida a respuestas coherentes + diversidad debida a respuestas incoherentes

o en términos más propios, varianza total = varianza verdadera + varianza debida a errores de medición

5. La fiabilidad la definimos como la proporción de varianza verdadera: *fiabilidad* = $\frac{\text{varianza verdadera}}{\text{varianza total}}$

En términos más simples:

$$\text{fiabilidad} = \frac{\text{varianza debida a respuestas coherentes (o relacionadas)}}{\text{varianza debida a respuestas coherentes y no coherentes}}$$

Decimos *respuestas distintas* porque suponemos que los sujetos son distintos, unos tienen *más* y otros *menos* del rasgo que medimos y decimos *repuestas coherentes* porque esperamos que cada sujeto responda de manera coherente (de manera parecida si todos los ítems expresan lo mismo).

6. El coeficiente de fiabilidad es un indicador de *relación global* entre las respuestas; expresa *cuánto* hay de relación en las respuestas. Esta relación es *relación verificada, empírica,* no es necesariamente *conceptual,* aunque la interpretación que se hace es *conceptual* (los ítems *miden lo mismo)*
Un coeficiente de, por ejemplo, .80 quiere decir que el 80% de la varianza se debe a *respuestas coherentes,* a lo que los ítems tienen *en común* o *de relacionado;* el 80% de la varianza total (de la *diversidad* que aparece en las puntuaciones totales) se debe a lo que los ítems tienen de relacionado.
7. La fiabilidad aumentará si aumenta el numerador, es decir 1º si hay diferencias en las respuestas y 2º si además las respuestas son coherentes (respuestas coherentes: las que *de hecho* están relacionadas).
8. Cómo se *interpreta* un coeficiente de fiabilidad *alto:*

   a) El test o escala *clasifica, ordena* bien a los sujetos en aquello que es *común a todos los ítems;*
   b) *Con un instrumento parecido encontraríamos resultados parecidos,* o si los sujetos respondieran muchas veces al mismo test o a tests semejantes, quedarían *ordenados* de manera similar (el coeficiente de fiabilidad es una *estimación* de la correlación esperable con un test paralelo).
   c) Los ítems *miden lo mismo* (por eso se llaman coeficientes de *consistencia interna);* generan *respuestas coherentes* y a la vez *distintas de sujeto a sujeto.* (Que los ítems *miden lo mismo hay que* interpretarlo con *cautela*; siempre es necesario un análisis conceptual y cualitativo).
9. La fiabilidad *tiende* a ser *mayor:*

   a) cuando los ítems expresan *lo mismo;* la definición del *rasgo* se expresa bien en todos los ítems;
   b) cuando es *mayor el número de ítems,* (con tal de que sean más o menos semejantes),
   c) cuando los ítems tienen un *mayor número de respuestas* (aunque no necesariamente),
   d) cuando los sujetos son más diferentes en aquello que se mide (*muestra heterogénea;* no se puede *clasificar* bien a los muy semejantes);
   e) en muestras *grandes* (porque hay más probabilidad de que haya sujetos más distintos).

14. Comentarios bibliográficos

1. La derivación de las fórmulas más conocidas del coeficiente de *fiabilidad* y otras relacionadas (como el *error típico,* etc.) pueden verse en Magnusson (1976). Entre las muchas obras que tratan de estos temas son especialmente recomendables las de Guilford (1954), Guilford y Fruchter, (1973), Nunnally (1978), Nunnally y Bernstein (1994), Thorndike (1982), Traub (1994). También disponemos de buenos artículos (Traub y Roley, 1991; Moss, 1994; Cronbach y Shavelson, 2004, del segundo autor utilizando notas de Cronbach fallecido en 1997, que resumen la historia de estos coeficientes).

2. La fórmula Kuder-Richardson 20 (y con más razón Kuder-Richardson 21, las dos más utilizadas con ítems dicotómicos) supone que todos los ítems tienen idéntica dificultad (media) e idéntica varianza; si esto no es así la fiabilidad resultante es una estimación más bien baja. Existen otros métodos que tienen en cuenta la diferente dificultad de los ítems, pero son más complicados; puede verse por ejemplo, en Horst (1953) y en Guilford y Fruchter (1973).

3. Ya hemos indicado que existen una serie de fórmulas de cálculo muy sencillo que simplifican las de Kuder-Richardson y otras como la del *error típico*. En general estas fórmulas no son recomendables dada la facilidad de cálculo que proporcionan calculadoras y ordenadores y además se trata solamente de *estimaciones* ya que suponen unas condiciones que no se suelen darse. Aun así pueden tener su utilidad para cálculos rápidos y aproximativos. Pueden encontrarse estas fórmulas en Saupe (1961) y en McMorris (1972), y para el *error típico* también en Burton (2004).

4. En las pruebas de rendimiento escolar no es siempre fácil dividir un test o prueba en dos mitades equivalentes para calcular la fiabilidad por el procedimiento de las dos mitades. También se puede calcular a partir de dos mitades de tamaño desigual o incluso a partir de tres partes (con muestras grandes en este caso). Se trata de procedimientos menos conocidos pero que pueden ser de utilidad en un momento dado; pueden encontrarse en Kristof (1974) y en Feldt (1975).

5. En los tests o pruebas objetivas de *criterio* (en los que hay una *puntuación de corte* para distinguir al *apto* del *no apto* y consecuentemente la distribución deja de ser normal) la fiabilidad se estima de otras maneras (pueden verse diversos índices en Mehrens y Lehmann, 1984, y en Berk, 1978); un índice apropiado y sencillo es el coeficiente de Livingston (puede verse en Mehrens y Lehmann, 1984; Black, 1999:291; en Black, 1999:292, tenemos también el *índice de discriminación* apropiado en los *tests de criterio).*

6. El coeficiente de fiabilidad también se puede calcular mediante *el análisis de varianza para muestras relacionadas,* con los mismos resultados que la fórmula del coeficiente α; puede verse en Hoyt (1941, 1952) y un ejemplo resuelto en Kerlinger (1975: 314-317) y en Fan y Thompson (2001). La relación entre fiabilidad y análisis de varianza también está explicada en Nunnally y Bernstein (1994: 274ss) y en Rosenthal y Rosnow (1991). Posiblemente como mejor se entiende la fiabilidad es desde el análisis de varianza.

7. Cómo calcular los *intervalos de confianza de los coeficientes de fiabilidad* puede verse en Fan y Thompson (2001); Duhachek y Iacobucci (2004) presentan tablas con el error típico de α para diversos valores del número de sujetos y de ítems y de la correlación media inter-ítem. El aportar estos intervalos de confianza es una de las recomendaciones *(guidelines)* de la *American Psychological Association* (Wilkinson and Task Force on Statistical Inference APA Board of Scientific Affairs, 1999).

8. Para verificar si dos coeficientes de fiabilidad (α) difieren significativamente puede verse Feldt y Kim (2006).

9. ***Fiabilidad inter-jueces.*** Un caso específico es el cálculo de la fiabilidad (o *grado de acuerdo)* entre diferentes evaluadores, cuando una serie de *jueces* evalúan una serie de sujetos, situaciones, etc. Puede utilizarse el *análisis de varianza para muestras relacionadas* que responde a esta pregunta: las diferencias observadas (la varianza total): ¿Se deben a que los jueces son distintos en su forma de evaluar, o a que los sujetos evaluados son distintos entre sí? De este análisis se deriva un

coeficiente que expresa lo mismo que el coeficiente α, pero la interpretación se hace sobre la *homogeneidad de los jueces* (o, con más propiedad, sobre el *grado de acuerdo* entre los jueces que aquí son los *ítems).* Este coeficiente da un valor muy parecido a la *correlación media* entre jueces (Rosenthal y Rosnow, 1991)[16].

Hay también otras *medidas de acuerdo entre jueces;* pueden verse, entre otros, en Holley y Lienert (1974) y Shrout y Fleiss (1979). El coeficiente *kappa* (κ) (Cohen, 1960) para medir el acuerdo entre *dos jueces* (datos dicotómicos, *unos* y *ceros;* κ = .60 se interpreta ya como un grado de *consensus* importante) es muy popular (puede encontrarse en numerosos textos, por ejemplo Fink, 1998; y sobre su interpretación Stemler, 2004). En Stemler (2004) pueden verse bien expuestos y valorados los diferentes enfoques para medir la fiabilidad de los jueces (*interrater reliability*).

16 La fiabilidad de los jueces calculada a partir del *análisis de varianza para muestras relacionadas* (disponible en EXCEL) es sencillo y de fácil comprensión por su relación con el coeficiente a de Cronbach; fórmula y explicación en Morales (2007).

15. Referencias bibliográficas

AMERICAN PSYCHOLOGICAL ASSOCIATION (2001). *Publication Manual of the American Psychological Association.* Washington D.C.:Author.

BERK, R.A. (1978). A consumers' guide to criterion-referenced tests item statistics. *NCME: Measurement in Education,* 9. 1

BLACK, THOMAS R. (1999). *Doing Quantitative Research in the Social Sciences.* London: Sage.

BURTON, RICHARD F. (2004). Multiple Choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*. 29 (5), 585-595.

CHARTER, RICHARD A. (2001). It Is Time to Bury the Spearman-Brown "Prophecy" Formula for Some Common Applications. *Educational and Psychological Measurement,* 61 (4). 690-696.

CRONBACH, LEE J. and SHAVELSON, RICHARD J. (2004). My Current Thoughts on Coefficient Alpha and Succesor Procedures. *Educational and Psychological Measurement,* 64 (3), 391-418.

COHEN, J., (1960), A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, 20, 1, 36-46.

DUHACHEK, ADAM and IACOBUCCI, DAWN (2004). Alpha's Standard Error (ASE): An Accurate and Precise Confidence Interval Estimate. *Journal of Applied Psychology***,** Vol. 89 Issue 5, p792-808.

FAN, XITAO and THOMPSON, BRUCE (2001). Confidence Intervals About Score Reliability Coefficients, please: An EPM Guidelines Editorial. *Educational and Psychological Measurement,* 61 (4), 517-531.

FELDT, LEONARD S., (1975), Estimation of the Reliability of a Test Divided into Two Parts of Unequal Length, *Psychometrika*, 40, 4, 557-561.

FELDT, LEONARD S. and KIM, SEONGHOON (2006). Testing the Difference Between Two Alpha Coefficients With Small Samples of Subjects and Raters. *Educational and Psychological Measurement*, 66 (4), 589-600

FINK, ARLENE (1998). *Conducting Research Literature Reviews, From Paper to the Internet.* Thousand Oaks & London: Sage Publications.

GARDNER, P. L. (1970). Test Length and the Standard Error of Measurement. *Journal of Educational Measurement* 7 (4), 271–273

GÓMEZ FERNÁNDEZ, D. (1981). El "ESP-E", un nuevo cuestionario de personalidad a disposición de la población infantil española. *Revista de Psicología General y Aplicada*, 36, 450-472.

GUILFORD, JOY P., (1954), *Psychometric Methods,* New York, McGraw-Hill;

GUILFORD, JOY P. and FRUCHTER, BENJAMIN, (1973), *Fundamental Statistics in Psychology and Education,* New York, McGraw-Hill.

HOLLEY, J.W. and LIENERT, G.A., (1974), The G Index of Agreement in Multiple Ratings, *Educational and Psychological Measurement,* 34, 817-822.

HORST, P., (1953), Correcting the Kuder-Richardson Reliability for Dispersion of Item Difficulties, *Psychological Bulletin,* 50, 371-374.

HOYT, C.J., (1941), Test Reliability Estimated by Analysis of Variance, *Psychometrika*, 3, 153-160.

HOYT, C.J., (1952), Estimation of Test Reliability for Un-Restricted Item Scoring Methods, *Educational and Psychological Measurement,* 12, 752-758.

KERLINGER, F. N., *Investigación del Comportamiento,* México, Interamericana.

KRISTOF, W. (1974), Estimation of the Reliability and True Score Variance from a Split of a Test into Three Arbitrary Parts, *Psychometrika,* 39, 4, 491-499.

MORALES VALLEJO, PEDRO (2007). *Análisis de varianza para muestras relacionadas.* www.upcomillas.es/personal/peter (última revisión, 26 de Agosto de 2007)

WILKINSON, LELAND and TASK FORCE ON STATISTICAL INFERENCE APA BOARD OF SCIENTIFIC AFFAIRS (1999*)* Statistical Methods in Psychology Journals: Guidelines and Explanations *American Psychologist* August 1999, Vol. 54, No. 8, 594–604 (Disponible en http://www.personal.kent.edu/~dfresco/CRM_Readings/Wilkinson_1999.pdf , consultado 22, 03, 2007)

MAGNUSSON, DAVID, (1976), *Teoría de los Tests,* México, Trillas.

MCMORRIS, R.F., (1972), Evidence of the Quality of Several Approximations for Commonly Used Measurement Statistics, *Journal of Educational Measurement,* 9, 2, 113-122.

MEHRENS, W. A. and LEHMANN, I.J. (1984). *Measurement and Evaluation in Education and Psychology* (3rd edition). New York: Holt, Rinehart and Winston.

MORALES VALLEJO, PEDRO (2006). *Medición de actitudes en Psicología y Educación*. 3ª edición. Madrid: Universidad Pontificia Comillas.

MORALES VALLEJO, PEDRO; UROSA SANZ, BELÉN y BLANCO BLANCO, ÁNGELES (2003). *Construcción de escalas de actitudes tipo Likert. Una guía práctica.* Madrid: La Muralla.

MOSS, PAMELA A., (1994), Can There Be Validity Without Reliability" *Educational Researcher,* 23, 2, 5-12

NUNNALLY, JUM C. and BERNSTEIN, IRA H. (1994). *Psychometric Theory*, 3rd. Ed. New York: McGraw-Hill.

NUNNALLY, JUM C., (1978), *Psychometric Theory,* New York, McGraw-Hill.

OSBORNE, JASON W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical Assessment, Research & Evaluation*, 8(11) http://PAREonline.net/getvn.asp?v=8&n=11

PFEIFFER, J.W.; HESLIN, R. AND JONES, J.E. (1976). *Instrumentation in Human Relations Training.* La Jolla, Ca.: University Associates.

ROSENTHAL, ROBERT AND ROSNOW, RALPH L. (1991). *Essentials of Behavioral Research, Methods and Data Analysis.* Boston: McGraw-Hill.

SAUPE, JOE L., (1961), Some Useful Estimates of the Kuder-Richardson formula number 20 Reliability Coefficient, *Educational and Psychological Measurement*, 21, 1, 63-71.

SCHMITT, NEAL (1996). Uses and abuses of Coefficient Alpha. *Psychological Assessment,* 8 (4), 350-353 (http://ist-socrates.berkeley.edu/~maccoun/PP279_Schmitt.pdf).

STREINER, DAVID L. (2003). Staring at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment,* 80 (1), 99-103.

SHROUT, PATRICK E. AND FLEISS, JOSEPH L., (1979), Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, 86, 420-428.

STEMLER, STEVEN E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4) http://pareonline.net/getvn.asp?v=9&n=4

THOMPSON, BRUCE (1994). Guidelines for authors. *Educational and Psychological Measurement,* 54, 837-847.

THORNDIKE, ROBERT L., (1982), *Applied Psychometrics*, Boston, Houghton Mifflin.

TRAUB, ROSS E., (1994), *Reliability for the Social Sciences: Theory and Applications*, Newbury Park, N. J., Sage.

TRAUB, ROSS E. AND ROWLEY, GLENN L., (1991), Understanding Reliability, *Educational Measurement: Issues and Practice*, 10 (1) 37-45.

WILKINSON, LELAND and TASK FORCE ON STATISTICAL INFERENCE *APA BOARD OF SCIENTIFIC AFFAIRS* (1999) Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist,* Vol. 54, No. 8, 594–604, http://www.spss.com/research/wilkinson/Publications/apasig.pdf (o en la página Web del autor http://www.spss.com/research/wilkinson/ en *online publications).*